在线学习资料支持

您可以在华为企业业务网站获得E-Learning课程、培训教材、产品资料、软件工具、技术案例等：

1、 **E-Learning课程：**登录*华为在线学习网站*，进入“*华为培训/在线学习*”栏目

免费E-Learning课：对网站所有用户免费开放

职业认证E-Learning课：通过任何一项职业认证即可学习所有职业认证培训E-Learning课程

渠道赋能E-Learning课：对华为企业业务合作伙伴免费开放

2、**培训教材：**登录*华为在线学习网站*，进入“*华为培训/面授培训*”，在具体课程页面即可下载教材。

华为职业认证培训教材、华为产品技术培训教材。无需注册即可下载

3、**华为在线公开课(LVC)：** http://support.huawei.com/ecommunity/bbs/10154479.html

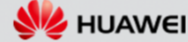
企业网络、UC&C、安全、存储等诸多领域的职业认证课程，华为讲师公开授课

4、**产品资料下载：** http://support.huawei.com/enterprise/#tabname=productsupport

5、**软件工具下载：** http://support.huawei.com/enterprise/#tabname=softwaredownload

更多内容请访问：

- http://learning.huawei.com/cn
- http://support.huawei.com/enterprise/
- http://support.huawei.com/ecommunity/



华为数通认证系列教程-HCDP IENP

提升企业级网络性能

Improving Enterprise Network Performance



华为技术有限公司

版权声明

华为数通认证系列教程-HCDP-Enterprise

华为认证数据通信资深工程师-企业级

第1.6版本

华为认证体系介绍

依托华为公司雄厚的技术实力和专业的培训体系，华为认证考虑到不同客户对ICT技术不同层次的需求，致力于为客户提供实战性、专业化的技术认证。

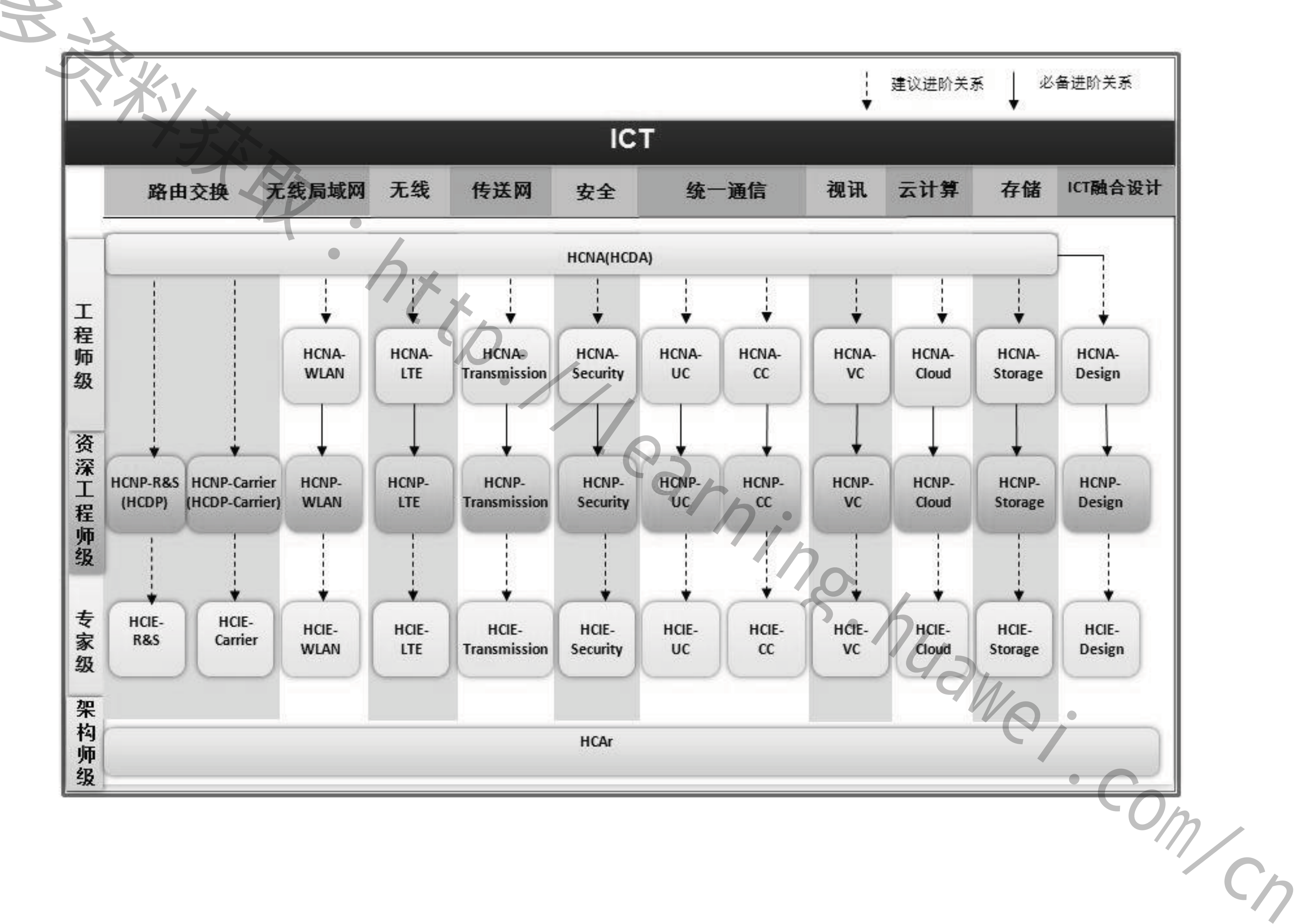
根据ICT技术的特点和客户不同层次的需求，华为认证为客户提供面向十三个方向的四级认证体系。

HCNA(HCDA)认证定位于中小型网络的基本配置和维护。HCNA(HCDA)认证包括但不限于：网络基础知识；流行网络的基本连接方法；基本的网络建造；基本的网络故障排除；华为路由交换设备的安装和调试。通过HCNA(HCDA)认证，将证明您对中小型网络有初步的了解，了解面向中小型企业的网络通用技术，并具备协助设计中小企业网络以及使用华为路由交换设备实施设计的能力。拥有通过HCNA(HCDA)认证的工程师，意味着中小企业有能力完成基本网络搭建，并将基本的语音、无线、云、安全和存储集成到网络之中，满足各种应用对网络的使用需求。

HCNP-Enterprise (HCDP-Enterprise)认证定位于中小型网络的构建和管理。HCNP-Enterprise (HCDP-Enterprise)认证包括但不限于：网络基础知识；交换机和路由器原理；TCP/IP 协议簇；路由协议；访问控制；网络故障的排除；华为路由交换设备的安装和调试。通过HCNP-Enterprise (HCDP-Enterprise)认证，将证明您对中小型网络有全面深入的了解，掌握面向中小型企业的网络通用技术，并具备独立设计中小企业网络以及使用华为路由交换设备实施设计的能力。拥有通过HCNP-Enterprise (HCDP-Enterprise)认证的工程师，意味着中小企业有能力完成完整网络的搭建，将企业中所需的语音、无线、云、安全和存储全面地集成到网络之中，并且能满足各种应用对网络的使用需求，进而提供较高的安全性、可用性和可靠性。

HCIE-Enterprise 认证定位于大中型复杂网络的构建、优化和管理。HCIE-Enterprise认证包括但不限于：不同网络和各种路由器交换机之间的互联；复杂连接问题的解决；使用技术解决方案提高带宽、缩短相应时间、最大限度地提高性能、加强安全性和支持全球应用；复杂网络的故障排除。通过HCIE-Enterprise认证，将证明您对大型网络有全面深入的了解，掌握面向大型企业网络的技术，并具备独立设计各种企业网络以及使用华为路由交换设备实施设计的能力。拥有通过HCIE-Enterprise认证的工程师，意味着大中企业有能力独立完成完整的网络搭建，将企业中所需的语音、无线、云、安全和存储全面地集成到网络之中，并且能满足各种应用对网络的使用需求；能够提供完整的故障排除能力；能根据企业和网络技术的发展，规划企业网络的发展，并提供高安全性、可用性和可靠性。

华为认证协助您打开行业之窗，开启改变之门，屹立在ICT世界的潮头浪尖！


建议进阶关系
必备进阶关系
ICT
路由交换
无线局域网
无线
传送网
安全
统一通信
视讯
云计算
存储
ICT融合设计
工程师级
资深工程师级
专家级
架构师级
HCNA(HCDA)
HCNA-WLAN
HCNA-LTE
HCNA-Transmission
HCNA-Security
HCNA-UC
HCNA-CC
HCNA-VC
HCNA-Cloud
HCNA-Storage
HCNA-Design
HCNP-R&S (HCDP)
HCNP-Carrier (HCDP-Carrier)
HCNP-WLAN
HCNP-LTE
HCNP-Transmission
HCNP-Security
HCNP-UC
HCNP-CC
HCNP-VC
HCNP-Cloud
HCNP-Storage
HCNP-Design
HCIE-R&S
HCIE-Carrier
HCIE-WLAN
HCIE-LTE
HCIE-Transmission
HCIE-Security
HCIE-UC
HCIE-CC
HCIE-VC
HCIE-Cloud
HCIE-Storage
HCIE-Design
HCAr

前言

简介

本书为 HCDP-IENP 认证培训教程，适用于准备参加 HCDP-IENP 考试的学员或者希望系统掌握华为安全产品与技术、可靠性 HA 技术、QoS 原理以及在华为通用路由平台 VRP 上的实现的读者。

内容描述

本书共包含三个 Module，系统地介绍了华为安全产品与技术、可靠性 HA 技术和 QoS 原理以及在 VRP 上的配置与实现。

Module1 详细介绍了华为 USG 防火墙产品功能特性和业务特性，使读
华为安全产品及网络安全有一个较为深入的了解；

Module 2 详细介绍了可靠性 HA 技术，帮助读者深入了解各种 HA 技术原理和运用。

Module 3 详细介绍了 IP QoS 技术，帮助读者深入了解 QoS 原理，掌握 QoS 在华为 VRP 中的配置。

本书引导读者循序渐进地掌握华为安全产品与技术、可靠性 HA 技术和 QoS 技术原理以及在华为产品中的实现，读者也可以根据自身情况选择感兴趣的章节阅读。

读者必备知识背景

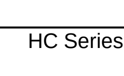
为了更好地掌握本书内容，阅读本书的读者应首先具备以下基本条件之一：

(1) 参加过 HCDA 培训

(2) 通过 HCDA 考试

(3) 熟悉 TCP/IP 协议，具有一定的网络基础知识

(4) 熟悉多种路由协议如 OSPF 和 BGP



本书常用图标



IPv6路由器



SOHO路由器



语音模块的路由器


中低端路由器


高端路由器


核心路由器


集线器


插座式交换机


汇聚交换机


核心交换机


边缘交换机


堆叠交换机


AP


AP大功率


无线网桥


无线网卡


接入服务器


语音网关


防火墙


网络电话系统

目 录

Module 1

华为安全产品与技术


USG防火墙产品基本
功能特性与配置
www.huawei.com
Copyright © 2012 Huawei Technologies Co., Ltd. All rights reserved.
HUAWEI



前　言

本胶片介绍了USG系列产品主要的安全技术和安全特性，以及各安全特性在USG产品上的配置。包括如：防火墙区域，防火墙工作模式，ASPF技术，NAT技术以及一些扩展技术。

Page1





培训目标

学完本课程后，您应该能：

- 掌握USG产品的主要安全技术和安全特性
- 掌握各安全特性在USG产品上的配置

HUAWEI



目 录

防火墙的基本概念

防火墙关键技术

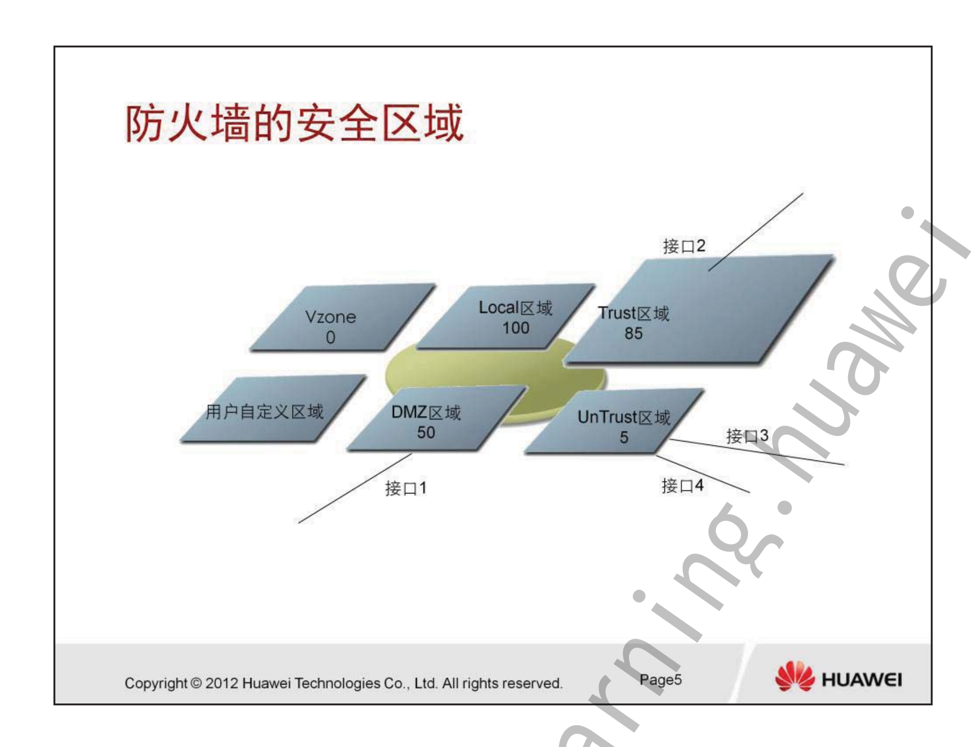
防火墙基本功能

防火墙扩展功能

Page3




目 录
1. 防火墙的基本概念
1.1 安全区域
1.2 防火墙工作模式
1.3 会话
Copyright © 2012 Huawei Technologies Co., Ltd. All rights reserved.
Page4
HUAWEI



域（Zone）是防火墙上引入的一个重要的逻辑概念；因为防火墙通常是放在网络的边界，路由器通过接口来连接不同网段，防火墙通过域来表示不同的网络，通过将接口加入域并在安全区域之间启动安全检查（称为安全策略），从而对流经不同安全区域的信息流进行安全过滤。常用的安全检查主要包括基于ACL和应用层状态的检查。

除Local区域外，使用其它安全区域前，都需要将安全区域分别与防火墙的特定接口关联，即将接口加入安全区域，接口只能加入到一个安全区域。该接口既可以是物理接口，也可以是逻辑接口。一个安全区域能够支持的最大接口数量为1024。

备注：接口添加进区域表达的意思是该接口所连接的网络属于该区域，但接口本身是属于Local区域的。

USG火墙上保留五个安全区域：

虚拟区(VZone)：虚拟防火墙所支持的区域，其安全优先级为0。

非受信区（Untrust）：低级的安全区域，其安全优先级为5。非军事化区（DMZ）：中度级别的安全区域，其安全优先级为50。

受信区（Trust）：较高级别的安全区域，其安全优先级为85。

本地区域（Local）：最高级别的安全区域，其安全优先级为100。

定义安全优先级的目的是用来区分安全区域间数据流的方向性，是inbound还是outbound，具体见下一页胶片。

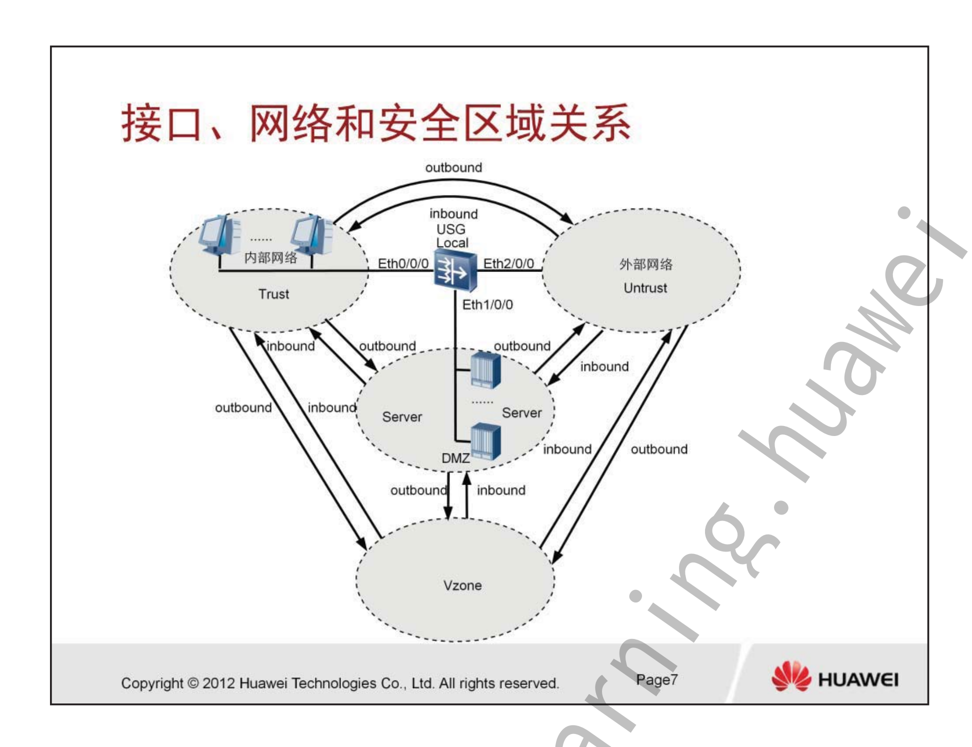
此外，如认为有必要，用户还可以自行设置新的安全区域并定义其安全优先级别。系统最多支持16个安全区域，包括五个保留的区域在内。

安全区域与各网络的关联遵循下面的原则：

内部网络应安排在安全级别较高的区域；

外部网络应安排在安全级别最低的区域；

一些可对外部提供有条件服务的网络应安排在安全级别中等的DMZ区。



当数据流在安全区域之间流动时，才会激发USG防火墙进行安全策略的检查，即USG防火墙的安全策略实施都是基于域间（例如Untrust区域和Trust区域之间）的，不同的区域之间可以设置不同的安全策略（例如包过滤策略、状态过滤策略等等）。

域间的数据流分两个方向：

- 入方向（inbound）:数据由低级别的安全区域向高级别的安全区域传输的方向；
- 出方向（outbound）：数据由高级别的安全区域向低级别的安全区域传输的方向。

备注：

- 任何两个安全区域的优先级不能相同；
- 本域内不同接口间的报文不过滤直接转发；
- 接口没有加入域之前不能转发报文。

在支持虚拟防火墙的情况下，有Vzone区域，Vzone区域也不包含任何接口，用于实现VPN实例间的报文转发。

数据流在VPN实例间的流动需要通过各自的Vzone区域跳转。

例如，当一个数据流从VPN1的Trust区域流向VPN2的DMZ区域时，该数据流需要首先从VPN1的Trust区域进入VPN1的Vzone区域；再从VPN2的Vzone区域进入VPN2的DMZ区域。其中，各VPN实例的Vzone区域是互相连通的，数据可以不受防火墙域间过滤规则的限制而自由流动。

安全区域配置－1

创建一个安全区域

[USG2100] firewall zone name *userzone*

设置优先级

[USG2100-zone-userzone] set priority 60

给安全区域添加接口

[USG2100-zone-userzone] add interface GigabitEthernet 0/0/1

Page9



firewall zone [name] zone-name

name：当创建一个新的安全区域或删除一个安全区域时，使用该关键字。进入保留的或已建立的安全区域视图时则不需要使用该关键字。

zone-name：安全区域名称。可取的字符集为：A~Z、a~z、0~9、“_”，必须以A~Z开头。名称对大小写不敏感，最大长度为32个字符。

用户可以自定义自己的安全区域，系统总共支持16个安全区域，包括系统默认的5个区域。自定义安全区域必须指定区域的优先级，通过set priority命令指定区域的优先级，优先级的范围为1－100，任何两个区域的优先级不能相同，并且系统默认的5个区域的优先级不能更改。

把接口添加进区域使用的命令是add interface，一个区域下最多支持的接口数量为1024个。



display zone [zone-name] [interface | priority]

zone-name：安全区域名称。

interface：显示隶属于安全区域的接口。

priority：显示安全区域的安全优先级。

命令display zone用来显示安全区域的配置信息，包括隶属于安全优先级、隶属于安全区域的接口等。

当不指定安全区域名称时则显示所有区域相关配置信息。

当不指定interface和priority关键字时，显示所有配置信息。

安全区域配置－2

在域间下发ACL

<USG2100>**system-view**

[USG2100] **policy interzone trust untrust outbound**

[USG2100-policy-interzone-trust-untrust-outbound] **policy 1**

[USG2100-policy-interzone-trust-untrust-outbound-1] **action permit**

 Page11



policy interzone trust untrust outbound :定义从trust到untrust区域的执行策略

inbound：过滤从低安全区域到高安全区域的数据包

outbound：过滤从高安全区域到低安全区域的数据包

policy 1 ：创建一个序号为1的策略

action permit：策略执行的结果是允许


目 录
1. 防火墙的基本概念
1.1 安全区域
1.2 防火墙工作模式
1.3 会话
Copyright © 2012 Huawei Technologies Co., Ltd. All rights reserved.
Page12
HUAWEI

防火墙的三种工作模式

路由模式

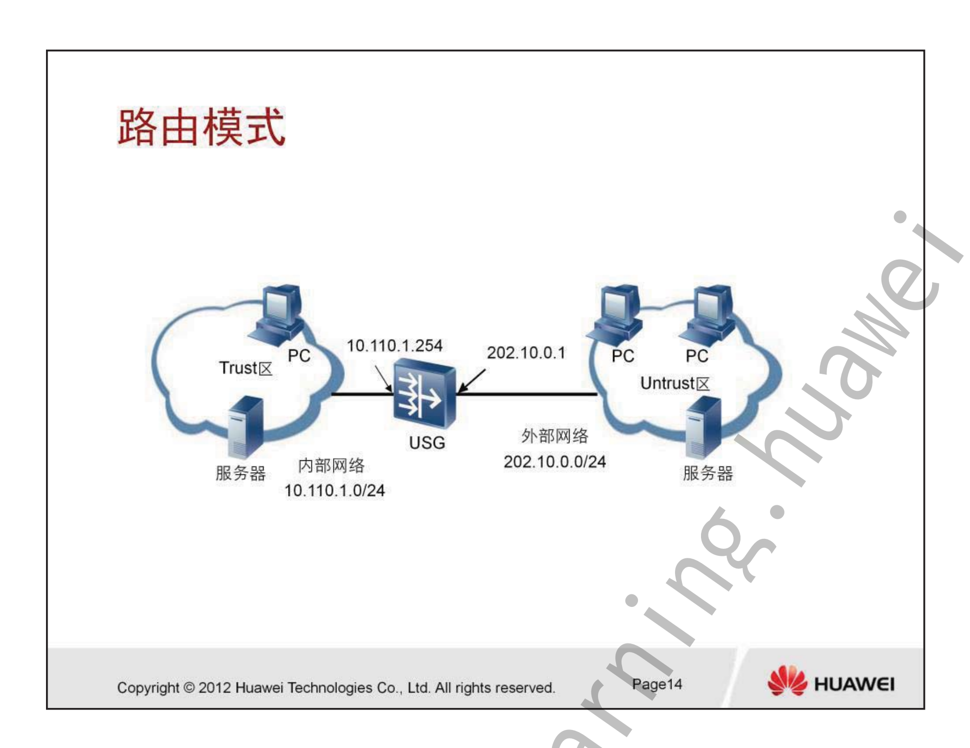
透明模式

混合模式

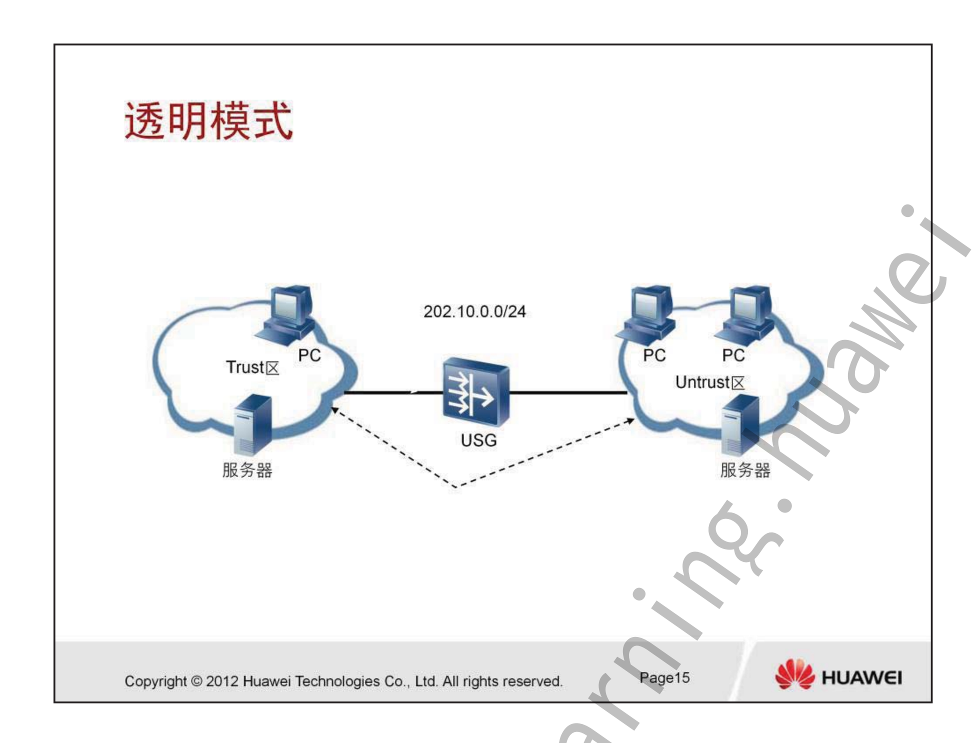
 Page13



依据USG防火墙的工作模式，具体可分为三种：路由模式，透明模式和混合模式。路由模式即防火墙像一台路由器，在防火墙接口上需要配置相关的IP地址，需要维护相关的路由信息。透明模式即防火墙像交换机等二层设备一样布置在网络之中，接口无需配IP地址，不改变网络的逻辑拓扑。混合模式即路由模式和透明模式的集会，防火墙的有些端口工作在三层，需要配置IP地址，有些端口工作在二层，像交换机端口，主要用于防火墙的双机备份应用。



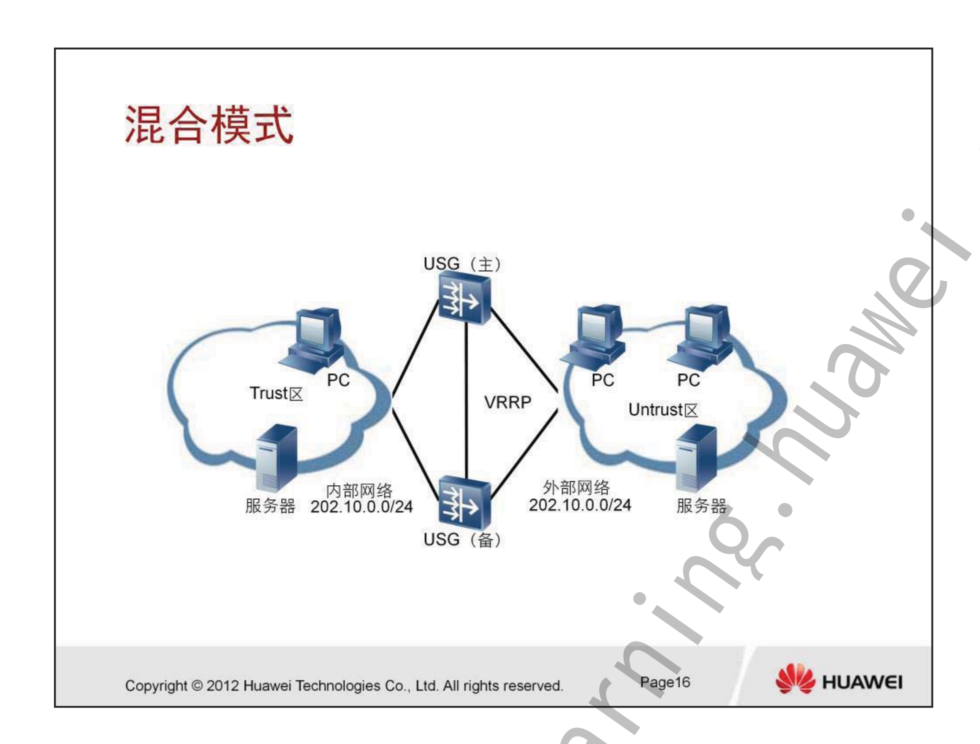
当USG防火墙位于内部网络和外部网络之间时，需要将防火墙与内部网络、外部网络以及DMZ三个区域相连的接口分别配置成不同网段的IP地址，重新规划原有的网络拓扑，此时防火墙相当于一台路由器。采用路由模式时，可以完成ACL包过滤、ASPF动态过滤、NAT转换等功能。然而，路由模式需要对网络拓扑进行修改。



如果USG防火墙采用透明模式进行工作，只需在网络中像放置网桥一样插入该USG防火墙设备即可，最大的优点是无需修改任何已有的配置；此时防火墙就象一个交换机一样工作，该工作模式在现网改造的时候很容易部署。

透明模式的防火墙支持ACL规则检查、ASPF状态过滤、防攻击检查、流量监控等功能。报文在防火墙当中不仅仅是像交换机的二层处理，会对报文进行高层分析处理。

防火墙上的地址表明，防火墙两端的接口处于同一个网段当中，就像交换机一样。防火墙接口不需要配置IP地址。



如果USG防火墙既存在工作在路由模式的接口（接口具有IP地址），又存在工作在透明模式的接口（接口无IP地址），则防火墙工作在混合模式下，这种工作模式基本上是透明模式和路由模式的混合，目前只用于透明模式下提供双机热备的特殊应用中，别的环境下不建议使用。

此时启动VRRP（Virtual Router Redundancy Protocol）功能的接口需要配置IP地址(即防火墙之间相连的接口)，其它接口不需要配置IP地址。此时内部网络和外部网络必须处于同一个子网中。


目 录
1. 防火墙的基本概念
1.1 安全区域
1.2 防火墙工作模式
1.3 会话
Copyright © 2012 Huawei Technologies Co., Ltd. All rights reserved.
Page17
HUAWEI

会话

会话（Session）

USG防火墙是状态防火墙，采用会话表维持通信状态。会话表包括五个元素：源IP地址、源端口、目的IP地址、目的端口和协议号(如果支持虚拟防火墙的话还有一个VPN-ID)。当防火墙收到报文后，根据上述五个元素查询会话表，并根据具体情况进行如下操作：

条件		操作
报文的五元组匹配会话表		转发该报文
报文的五元组不匹配会话表	域间规则允许通过	转发该报文，并创建会话表表项
	域间规则不允许通过	丢弃该报文

Page18



会话是状态防火墙的基础，每一个通过防火墙的会话都会在防火墙上建立一个会话表项，以五元组（源目的IP地址、源目的端口、协议号）为Key值，通过建立动态的会话表来为高优先级域提供更高的安全性。



display firewall session table [verbose]

用来现实系统当前的会话表项信息，verbose参数来控制是否显示详细的信息。

如上所示：

Current Total Sessions：1当前会话表数统计。

- telnet 说明会话类型是使用telnet协议。
- VPN 表示方式为：源方向-->目的方向。

TTL 该会话表项总的生存时间。

Left 该会话表项剩余生存时间。

Interface 出接口。

NextHop 下一跳IP地址。

MAC 下一跳MAC地址。

<--packets:1269 bytes:66769 该会话入方向的包数和字节数统计。

-->packets:1081 bytes:43715 该会话出方向的包数和字节数统计。

Reset Session表项操作得谨慎，因为在运行业务将会中断。



有了动态创建的会话表后，会话表就成为了一个资源，为避免资源耗尽，防火墙上的会话表都有老化时间，根据应用的不同可以单独设定；在指定的时间内没有报文通过的话会话表就会被老化掉。

display firewall session aging-time 用来显示所有的表项老化时间的设置。该值对应命令display firewall session table verbose显示的表项中的TTL值。

系统对于各种协议都有默认的表项老化时间，单位为秒。

当然，用户可以根据需要修改这些默认的值。比如把ICMP类型的老化时间改成15秒。



会话表项的老化机制在一些特殊的应用中会导致问题，因为实际应用中，防火墙某些会话类型的老化时间相对较短，而对应数据流很长时间没有报文刷新，但是Session需要保持不能被老化。为解决这一矛盾，防火墙提出长连接特性。长连接功能用于设置特定数据流的超长老化时间。该数据流由ACL确定。数据流的老化时间不受全局老化时间限制。

防火墙使用命令firewall long-link aging-time来调整表项的老化时间。

firewall long-link aging-time interval

interval：长连接老化时间。范围为1h～480h，缺省值为168h。

long-link acl-number { inbound | outbound }

acl-number：ACL规则号。范围为3000～3999。

inbound：在安全域间的入方向使能长连接功能。

outbound：在安全域间的出方向使能长连接功能。

该命令用于使能防火墙的长连接功能。域间的入域方向和出域方向可以单独或同时关联ACL规则。入域方向和出域方向可以关联不同的ACL规则。

配置过程中，建议不要引用范围过大的ACL规则，否则会影响到防火墙的性能。



目 录

Page22




目 录
2. 防火墙关键技术
2.1 ASPF
Copyright © 2012 Huawei Technologies Co., Ltd. All rights reserved.
Page23
HUAWEI

ASPF

ASPF（Application Specific Packet Filter）是一种改进的高级通信过滤技术，ASPF不但对报文的网络层的信息进行检测，还能对丰富的应用层协议进行深度检测，支持多媒体业务的NAT以及安全防范功能，支持的协议包括：H.323协议族、MGCP、SIP、H248、RTSP、HWCC及ICMP、FTP、DNS、PPTP、NBT、ILS、HTTP、SMTP等。

基于ACL规则的包过滤可以在网络层和传输层检测数据包，防止非法入侵。

ASPF对应用层的协议信息进行检测，通过维护会话的状态和检查会话报文的协议和端口号等信息，阻止恶意的入侵。

 Page24 

ASPF在Session表的数据结构中维护着连接的状态信息，并利用这些信息来维护会话的访问规则。ASPF保存着不能由访问列表规则保存的重要的状态信息。防火墙检验数据流中的每一个报文，确保报文的状态与报文本身符合用户所定义的安全规则。连接状态信息用于智能的允许/禁止报文。当一个会话终止时，Session表项也将被删除，防火墙中的会话也将被关闭。

对于TCP连接，ASPF可以智能的检测“TCP的三次握手的信息”和“拆除连接的握手信息”，通过检测握手、拆连接的状态检测，保证一个正常的TCP访问可以正常进行，而对于非完整的TCP握手连接的报文会直接拒绝。

UDP是无连接的报文，所以也没有真正的UDP“连接”。因为ASPF是基于连接的，它将对UDP报文的源、目的IP地址、端口进行检查，通过判断该报文是否与所设定的时间段内的其他UDP报文相类似，而近似判断是否存在一个连接。

在普通的场合，一般使用的是基于ACL的IP包过滤技术，这种技术比较简单，但缺乏一定的灵活性，在很多负责应用的场合普通包过滤是无法完成对网络的安全保护的。例如对于类似于应用FTP协议进行通信的多通道协议来说，配置防火墙则是非常困难的。

ASPF使得USG系列防火墙能够支持一个控制连接上存在多个数据连接的协议，同时还可以在应用非常复杂的情况下方便的制订各种安全的策略。大部分多媒体应用协议（如H.323、SIP）及FTP、Net Meeting等协议使用约定的端口来初始化一个控制连接，再动态的选择端口用于数据传输。端口的选择是不可预测的，其中的某些应用甚至可能要同时用到多个端口。因此包过滤防火墙只有阻止单通道的应用传输，以免内部网络遭受攻击，只能仅仅阻止了一些使用固定端口的应用，而留下了许多安全隐患。而ASPF监听每一个应用的每一个连接所使用的端口，打开合适的通道让会话中的数据能够出入防火墙，在会话结束时关闭该通道，从而能够对使用动态端口的应用实施有效的访问控制。

多通道协议

多通道协议

- 是指某个应用在进行通讯或提供服务时需要建立两个以上的会话（通道），其中有一个控制通道，其他的通道是根据控制通道中双方协商的信息动态创建的，一般我们称之为数据通道或子通道。多通道协议在状态防火墙当中需要特殊处理。

单通道协议

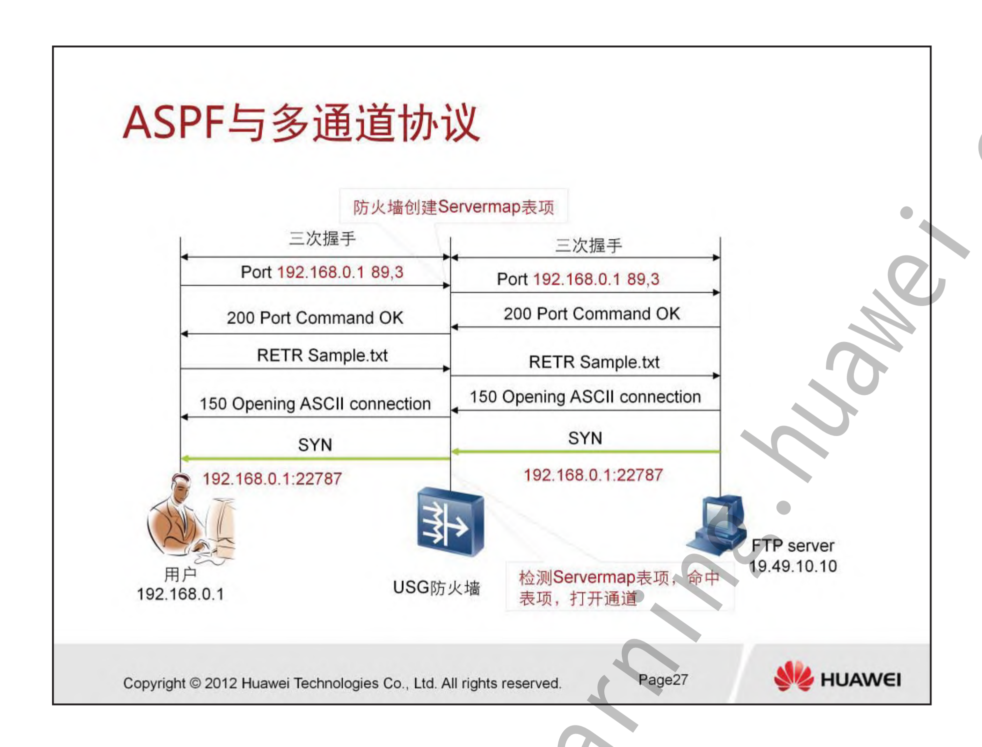
- 是指某个应用在进行通讯或提供服务时只需要建立一个会话的应用协议。根据TCP三次握手机制，状态防火墙能够维护会话的五元组信息。

Page26



多通道协议在防火墙应用以及NAT设备的应用中需要特殊处理，因为数据通道的端口是不固定的（协商出来的）其报文方向也是不固定的。基于防火墙的安全策略，没有开放的端口报文是会被拒绝的，所以对于动态协商的端口，防火墙需要动态的打开，利用ASPF技术。

多通道协议的典型应用有很多，最典型的是FTP，其他的一般都和音频或视频通讯有关如：H.323（包括T.120、RAS、Q.931和H.245等）、SIP、MGCP，此外许多实时聊天通讯软件也是多通道协议的，如MSN，QQ，ICQ等。



FTP应用包含一个预知端口（21）的TCP控制通道和一个动态协商的TCP数据通道，对于一般的包过滤防火墙来说，配置安全策略时无法预知数据通道的端口号，因此无法确定数据通道的入口。这样就无法配置准确的安全策略。ASPF技术则解决了这一问题，它检测IP层之上的应用层报文信息，并动态地根据报文的内容创建和删除临时的Servermap表项，以允许相关的报文通过。

从上图中可以看出，Servermap表项是对FTP控制通道中数据动态检测的过程中动态产生的，当报文通过防火墙时，ASPF将对报文与指定的访问规则进行比较，如果规则允许，报文将接受检查，否则报文直接被丢弃。如果该报文是用于打开一个新的控制或数据连接，ASPF将动态的产生Servermap表项，对于回来的报文只有是属于一个已经存在的有效的连接，才会被允许通过防火墙。在处理回来的报文时，状态表也需要更新。当一个连接被关闭或超时后，该连接对应的状态表将被删除，确保未经授权的报文不能随便透过防火墙。因此通过ASPF技术可以保证在复杂应用的情况下，依然可以非常精确的保证网络的安全。

USG防火墙在对多通道协议支持的时候，采用了5元组的Session表项＋3元组的Servermap表项的结合方式实现，这样在很大程度上保证了内部网络的安全性。因为Servermap表项是一个“临时入口”表项，

当真正的数据报文来了以后，会根据这个数据报文＋Servermap进行完整判断，确定这个连接是一个合法的数据通道。当数据通道建立了之后，会为这个数据通道建立一个基于5元组的Session通道，而Servermap表项不再有报文命中，就会被老化掉，这样就避免了因为多通道协议的特点，而产生一个永久的三元组通道，避免了内部网络暴露的安全隐患。

ASPF检测主控制通道数据应用层上的内容以获得相关参数。数据通道建立后，对数据通道不做内容检测。

对于上图中，端口号的计算方式为 22787＝89*256+3,这是FTP的报文格式所约定的。

三元组ASPF

USG相当于一个六元组(支持VPN情况下，有VPN-ID)的NAT设备，即防火墙上的每个会话的建立都需要六元组：源IP地址、源端口、目的IP地址、目的端口、协议号和VPN-ID。只有这些元素都具备了，会话才能建立成功，报文才能通过。而一些实时通讯工具，如QQ、MSN等，通过NAT设备，需要按三元组处理：源IP地址、源端口、协议号。 USG为了适配类似QQ、MSN等通讯机制，支持三元组处理方式，让类似QQ、MSN等的通讯方式能够正常的穿越。

除QQ、MSN穿越NAT设备外，其他仅使用源IP地址、源端口、协议号的会话，如TFTP，同样需要配置防火墙三元组ASPF。

 Page29 

TFTP是一个传输文件的简单协议，它其于UDP协议而实现，但是我们也不能确定有些TFTP协议是基于其它传输协议完成的。此协议设计的时候是进行小文件传输的。因此它不具备通常的FTP的许多功能，它只能从文件服务器上获得或写入文件，不能列出目录，不进行认证，它传输8位数据。传输中有三种模式：netascii，这是8位的ASCII码形式，另一种是octet，这是8位源数据类型；最后一种mail已经不再支持，它将返回的数据直接返回给用户而不是保存为文件。

因为TFTP基于UDP，而UDP基于IP，IP可以还使用其它本地通信方法。因此一个TFTP包中会有以下几段：本地媒介头，IP头，UDP头，TFTP头，剩下的就是TFTP数据了。TFTP在IP头中不指定任何数据，但是它使用UDP中的源和目标端口以及包长度域。由TFTP使用的包标记（TID）在这里被用做端口，因此TID必须介于0到65,535之间。对它的初始化我们在后面讨论。TFTP头中包括两上字节的操作码，这个码指出了包的类型下面我们看看大体上的TFTP包格式。

| Local Medium | Internet | Datagram | TFTP |

初始连接时候需要发出WRQ（请求写入远程系统）或RRQ（请求读取远程系统），收到一个确定应答，一个确定可以写出的包或应该读取的第一块数据。通常确认包包括要确认的包的包号，每个数据包都与一个块号相对应，块号从1开始而且是连续的。因此对于写入请求的确定是一个比较特殊的情况，因此它的包的包号是0。如果收到的包是一个错误的包，则这个请求被拒绝。创建连接时，通信双方随机选择一个TID，因为是随机选择的，因此两次选择同一个ID的可能性就很小了。每个包包括两个TID，发送者ID和接收者ID。这些ID用于在UDP通信时选择端口，请求主机选择ID的方法上面已经说过了，在第一次请求的时候它会将请求发到TID 69，也就是服务器的69端口上。应答时，服务器使用一个选择好的TID作为源TID，并用上一个包中的TID作为目的ID进行发送。这两个被选择的ID在随后的通信中会被一直使用。

因为TFTP没有安全控制机制，因此安全问题应该多加考虑。通常TFTP允许下载数据而不允许上传数据。

ASPF配置

进入安全区域域间

[USG2100] firewall interzone trust untrust

打开ASPF功能

[USG2100-interzone-trust-untrust] detect *protocol*

[USG2100-interzone-trust-untrust] detect { **activex-blocking** | **java-blocking** }

Page31



Detect protocol：ASPF支持的协议名称。可选参数可以为：

FTP、H.323、HTTP、HWCC（Huawei Conference control Protocol）、MSN、NetBIOS、PPTP、QQ、RTSP。

protocol参数还可以为all、activex-blocking和java-blocking，分别表示设置所有ASPF应用协议检测以及设置ActiveX阻断、Java小程序阻断时的ACL访问控制列表。

Java Blocking（Java阻断）：保护网络不受有害Java Applets的破坏。

ActiveX Blocking（ActiveX阻断）：保护网络不受有害ActiveX的破坏。

inbound：过滤从接口收上来的数据包。

outbound：过滤从接口转发的数据包。

备注：

使用detect all命令不能设置Java阻断和ActiveX阻断；

使用undo detect all命令不能取消所设置的Java阻断和ActiveX阻断。



目 录

HUAWEI



 Page33 HUAWEI

黑名单

黑名单特点:

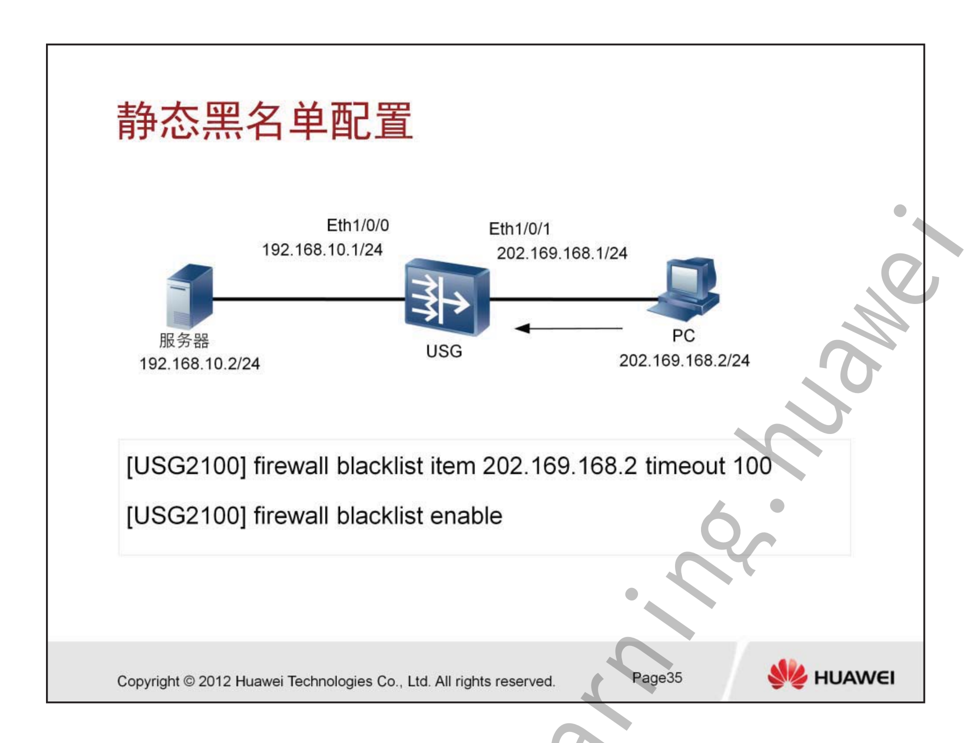
- 根据报文的源IP地址进行过滤
- 简单高效
- 可动态添加删除

Page34



黑名单指根据报文的源IP地址进行过滤的一种方式。同基于ACL的包过滤功能相比，由于黑名单进行匹配的域非常简单，可以以很高的速度实现报文的过滤，从而有效地将特定IP地址发送来的报文屏蔽。黑名单最主要的一个特色是可以由USG防火墙动态地进行添加或删除，当防火墙中根据报文的行为特征察觉到特定IP地址的攻击企图之后，通过主动修改黑名单列表从而将该IP地址发送的报文过滤掉。因此，黑名单是防火墙一个重要的安全特性。

黑名单因为其动态特性，规则添加简单以及对报文过滤的操作简单性而曾经被广泛应用，不过随着高速ACL相关技术被广泛应用而逐渐失去优势，特别是在一些应用场景会出现一些问题，因此在实际应用中不推荐使用。使用的时候一定要小心。

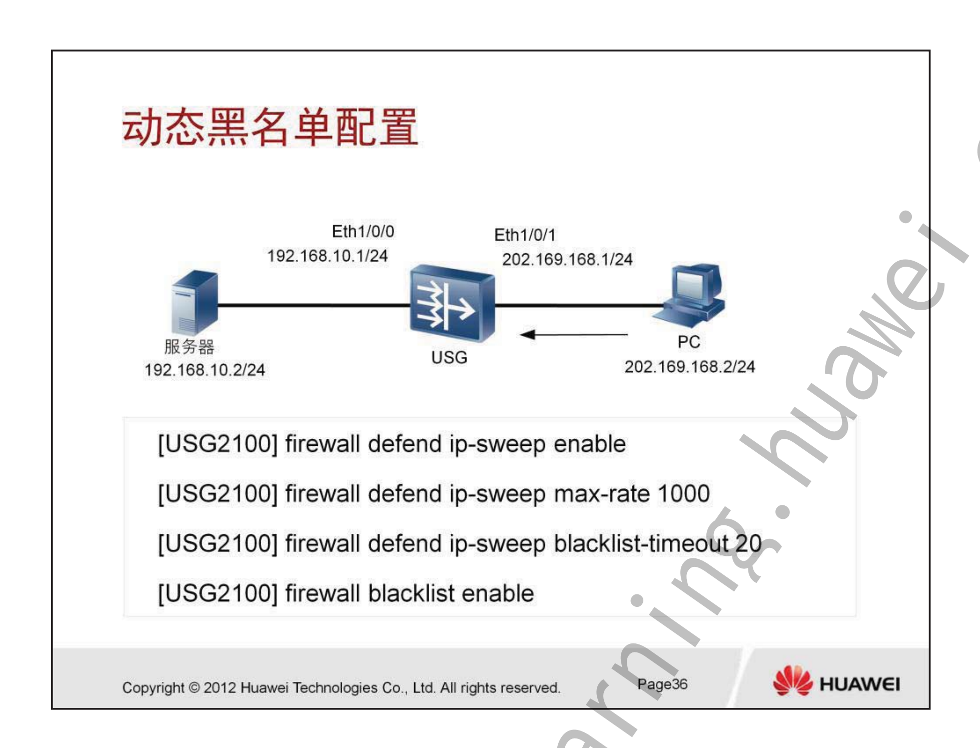


firewall blacklist item source-address [timeout interval]

若使用timeout interval参数，则意味着该黑名单表项会在指定的老化时间后被自动删除，从而对源自相应IP地址的报文过滤功能也随之消失。

若不使用该参数，则意味着该表项永远有效，不会被老化。

timeout 的单位为分钟，取值范围为1-1000。



动态黑名单配置，和防火墙的攻防模块配合：

使能地址扫描攻击防范功能。

[USG2100] firewall defend ip-sweep enable

配置抵制扫描攻击防范功能：

[USG2100] firewall defend ip-sweep max-rate 1000　　//从同一源地址向外发送报文的目的IP变化速率的阈值，如果变更速度超过1000个每秒，则认为是IP扫描攻击；取值范围为1次/秒～10,000次/秒，默认4000次/秒。

[USG2100] firewall defend ip-sweep blacklist-timeout 20 //将攻击源IP加入黑名单并设定其在黑名单内的保持时间，取值范围为1min～1000min，默认值为20min，即不加入黑名单。

[USG2100] firewall blacklist enable　　//命令用来开启黑名单功能。acl-number ACL的编号，整数形式，取值范围为2000～3999。

黑名单配置验证

```
[USG2100] display firewall blacklist item
Total:1        Manual:1          IP Sweep:0        Port Scan:0
 IDS:0          Login Failed:0     PreAuthed:0       Get Flood:0
 tcp-illeage-session:0             Unknown:0
 IP                 Reason     InsertTime                 AgeTime  Vpn-
    instance
 -----------------------------------------------------------------------------------------
202.169.168.2   Manual    2009/05/12 17:47:35  Permanent
```

 Page37 HUAWEI

display firewall blacklist { enable | item [source-address] }

参数：

enable：显示黑名单功能运行状况。

item source-address：显示黑名单表项内容。source-address为表项的IP地址。

描述：

命令display firewall blacklist显示防火墙黑名单功能的运行状况和黑名单表项信息。参数item [sour-address]显示黑名单表项信息，如果不指定IP地址，显示当前全部黑名单表项的概要信息，如果指定了要显示的IP地址，显示特定黑名单表项的详细信息。参数enable显示黑名单功能的运行状况。

Age Time表示黑名单的老化时间；

Insert-Time表示表项已经经过的时间，逐渐递增至老化时间。

REASON：Manual表示为静态黑名单表项；Logging failed表示Telnet三次登录失败的黑名单表项；IPSweep/PortScan表示IP地址扫描或端口扫描黑名单表项。



目 录

3. 防火墙基本功能

Page38



MAC绑定

问题的提出

- 网络中常有一些假冒IP地址的攻击

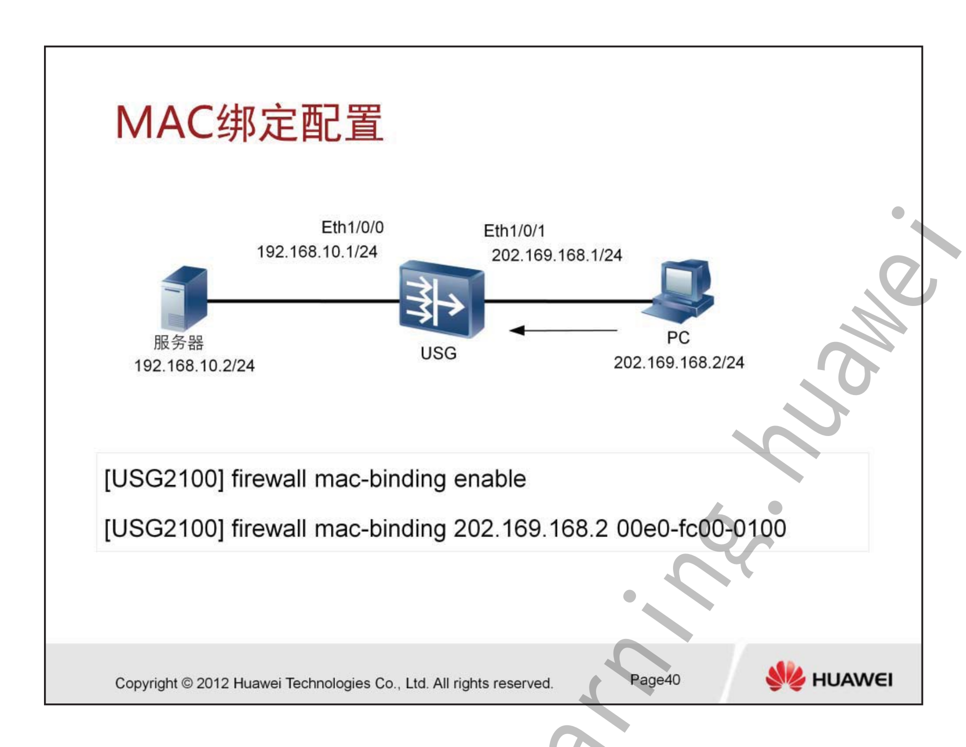
MAC绑定应用限制条件

- 与二层直接相连的网络

 Page39 

MAC和IP地址绑定，指防火墙可以根据用户的配置，在特定的IP地址和MAC地址之间形成关联关系。对于声称从这个IP发送的报文，如果其MAC地址不是指定关系对中的地址，防火墙将予以丢弃。发送给这个IP地址的报文，在通过防火墙时将被强制发送给绑定的MAC地址，从而形成有效的保护，是避免IP地址假冒攻击的一种方式。

MAC和IP地址绑定功能一般应用在与二层交换机直接相连的时候，可以防止假冒IP地址攻击，ARP Flood攻击，DHCP Flood攻击等，还可以应用于用户认证。



firewall mac-binding { enable | ip-address mac-address }

参数:

enable：使能地址绑定功能，配置Mac绑定一定要使能绑定功能，否则绑定不生效。

ip-address：指定地址绑定对的IP地址。

mac-address：指定地址绑定对的MAC地址。

MAC绑定配置验证

```
[USG2100] display firewall mac-binding item
 Firewall Mac-binding items :
 Current items : 3
 192.168.2.18       0087-0326-ea9d
 202.1.1.8          00e0-fc08-0589
 202.1.1.9          00e0-fc98-5679
```

 Page41 

display firewall mac-binding { enable | item [ip-address] }

参数：

enable：显示地址绑定功能运行状况。

item：显示地址绑定表项内容。

ip-address：要显示的表项的IP地址。

描述：

命令display firewall mac-binding显示防火墙地址绑定功能的运行状况和地址绑定表项信息。参数item [ip-address]显示地址绑定表项信息，如果不指定IP地址，显示当前全部地址绑定表项的概要信息，如果指定了要显示的IP地址，显示特定地址绑定表项的详细信息。参数enable显示地址绑定功能的运行状况。



目 录

 Page42 HUAWEI

端口映射

问题的提出

- 内部服务器在非知名端口提供知名服务，例如在1021端口提供FTP服务

端口映射

- 防火墙并非要更改数据包的端口信息
- 可以用来保护因为知名端口而带来的针对性攻击

 Page43 

端口映射是解决端口和服务类型映射关系的一个配置功能，对于使用非知名端口提供知名服务时的业务识别非常有用，典型的ftp的服务端口是21，如果有些用户将1021端口作为ftp的服务端口，怎么办呢？可以通过port-mapping命令来绑定该对应关系，这样1021端口的服务就被自动识别成ftp服务了，另外如果该端口只对特定地址有效可以通过ACL来限制识别范围。

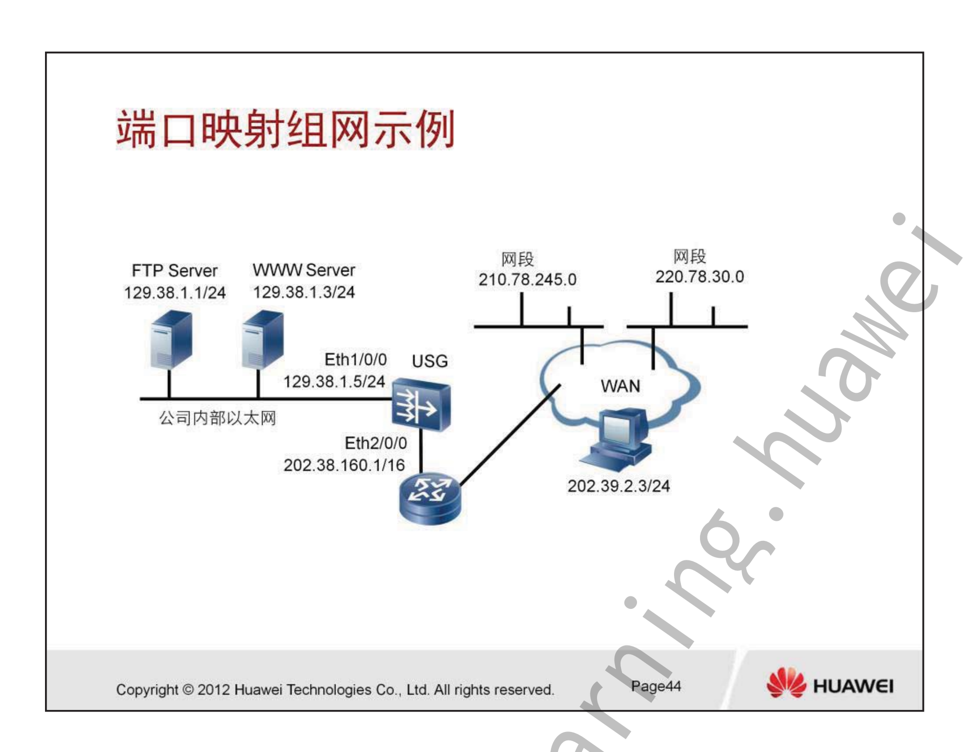
端口映射能够对不同的应用协议创建和维护一张系统定义（system-defined）和用户定义（user-defined）的端口识别表。

USG防火墙支持基于基本ACL的主机端口识别。这种主机端口识别是对去往某些特定主机的报文建立自定义端口号和应用协议的识别。例如：将去往10.110.0.0网段的主机使用8080端口的TCP报文识别为HTTP报文。主机的范围可由基本ACL指定。

需要指出，主机端口识别与包过滤引用的ACL有如下差别：

包过滤中，防火墙只允许从“源地址”到“目的地址”的报文通过；

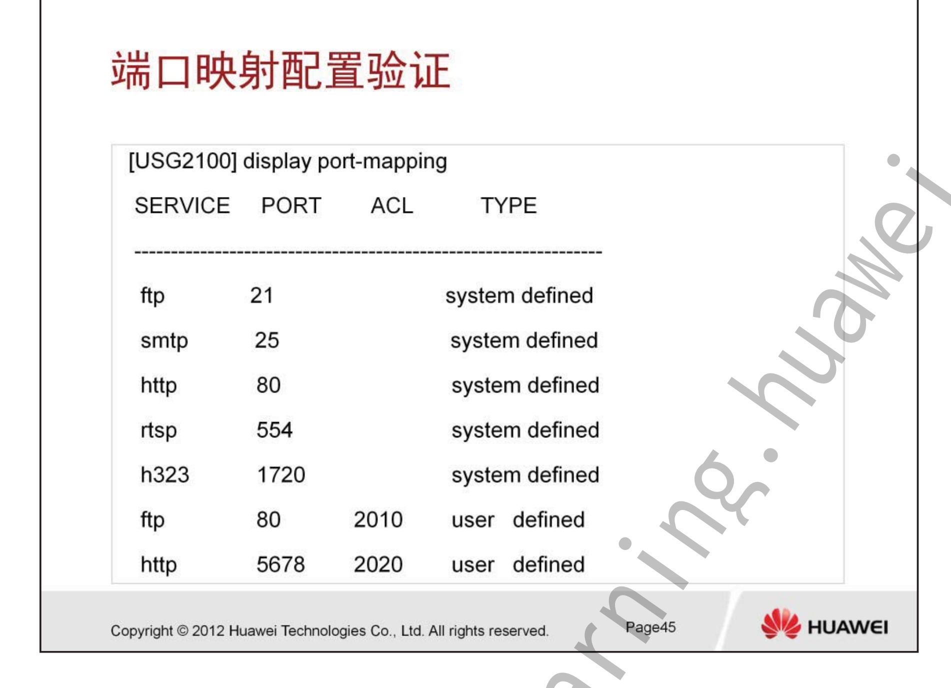
主机端口识别中，只用基本ACL定义主机的范围即可，不存在方向性，即源或者目的IP匹配该ACL定义的范围即认为匹配。



如上图所示，某公司对外提供WWW和FTP服务，通过配置防火墙，希望能够识别以下端口：

将去往主机129.38.1.1/32的使用端口号80的报文识别为FTP报文；

将去往网段129.38.1.0/24的使用端口号5678的报文识别为HTTP报文。



display port-mapping [protocol-name | port port-number]

参数：

protocol-name：指定用于端口识别的应用的名称。可选的应用有：ftp、http、h323、smtp、rtsp。

port port-number：端口识别的端口号，port-number取值范围是0～65535。



port-mapping protocol-name port port-number acl acl-number

参数:

protocol-name：应用的名称。可选的应用有：ftp、http、h323、smtp、rtsp。

port-number：端口号，取值范围是0～65535。

acl-number：基本访问列表号，取值范围是2000～2999。

描述:

命令port-mapping用来建立端口到应用层协议的识别。

USG2100防火墙支持基于基本访问控制列表（ACL）的主机端口识别。基于基本访问控制列表的主机端口识别是对去往某些特定主机的报文建立自定义端口号和应用协议的识别，主机的范围由基本的ACL指定。



目 录

Page47



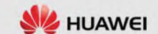
IDS联动

防火墙的局限性

- 防火墙不能防止通向站点的后门；
- 防火墙一般不提供对内部的保护；
- 防火墙无法防范数据驱动型的攻击；
- 防火墙不能根据网络被恶意使用和攻击的情况动态调整自己的策略等。

IDS（Intrusion Detection System，入侵检测系统）的优势

- 实时地监视、分析网络中所有的数据报文，发现并实时处理所捕获的数据报文，对系统记录的网络事件进行统计分析，发现异常现象，主动切断连接或与防火墙联动，调用其他程序处理。

 Page48 

由于防火墙自身具有一定的局限性，如检查的颗粒度较粗，难以对众多的协议细节进行深入的分析与检查，并且防火墙具有防外不防内的特点，难以对内部用户的非法行为和已经渗透的攻击进行有效的检查和防范，对数据驱动型攻击无法防范等。因此，USG防火墙开放了相关接口，通过与其它安全软件进行联动，从而构建统一的安全网络。

IDS入侵检测系统的作用：

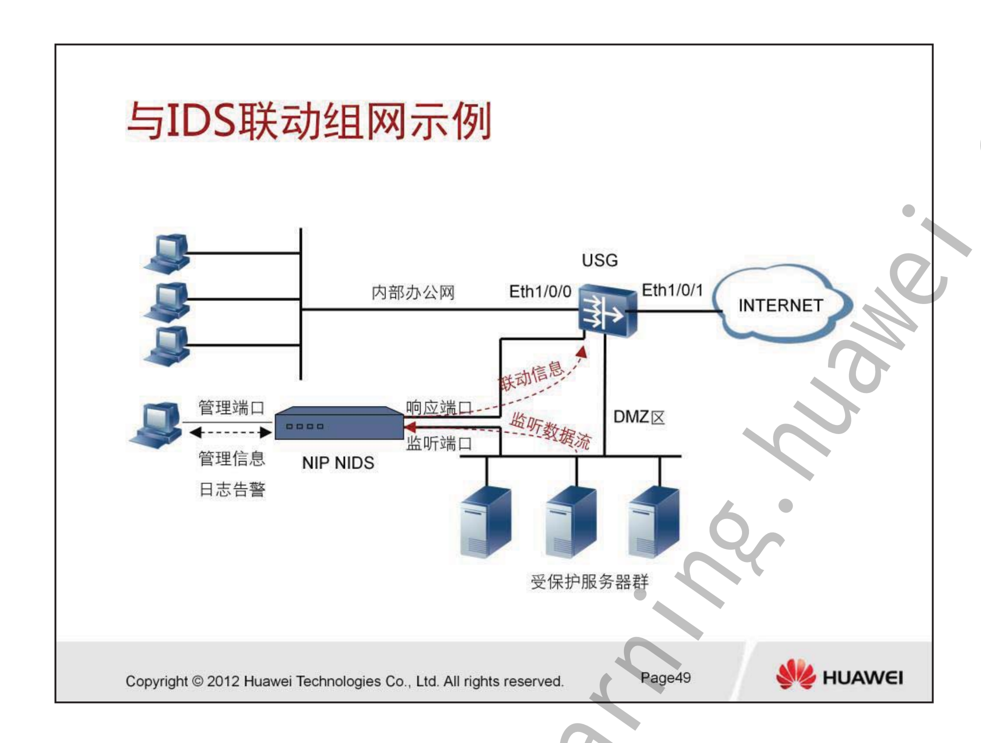
- 实时地监视、分析网络中所有的数据报文
- 发现并实时处理所捕获的数据报文
- 对系统记录的网络事件进行统计分析
- 发现异常现象，主动切断连接或与防火墙联动，调用其他程序处理

IDS系统就像在网络上装备了网络分析器，对网络传输进行监视。该系统熟悉最新的攻击手段，而且尽力检查通过的每个报文，从而尽早处理可疑的网络传输。具体采取的措施由用户使用的特定IDS系统和配置情况决定。

USG防火墙可以联动的IDS系统有：

启明星辰的IDS；安氏IDS；天龙马IDS；金诺网安IDS；华为NIP IDS





防火墙定义相关的安全区域，如上图，IDS监测进入DMZ区域的数据。通过端口镜像技术，把DMZ区域的流量镜像一份给NIP IDS，IDS基于此数据进行检测，一旦发现有异常情况，通过报警通知系统管理员和与防火墙联动，由防火墙来阻断相应的攻击流。



firewall ids server ip-address

firewall ids server命令用于配置外接IDS服务器的IP地址。

firewall ids port port-number

port port-number：设置使用的端口号，取值范围是1025～65535，缺省为40000，防火墙通过40000端口与外接IDS服务器通讯 。

firewall ids authentication type { md5 [key key-string] | none }

md5：与外接IDS服务器采用MD5进行报文认证。

none：与外接IDS服务器不进行报文认证。

key key-string：未加密的密钥，取值范围是字符串，长度为1～16。

缺省情况下，与外接IDS服务器不进行报文认证，即采用none方式。

firewall ids enable命令用于启动外接IDS功能。

打开防火墙外接IDS功能时，应首先配置IDS服务器的IP地址以及报文的认证方式。

IDS配置验证

```
[USG2100] display firewall ids
Firewall IDS information:
  firewall IDS: enable
  debug flag: off
  server port: 3000
  authentication type: md5
  authentication string: huawei123
  client address 0: 192.168.10.10
```

 Page51 

debug flag: off

如果防火墙上打开了IDS的调试开关，则debug flag为on。

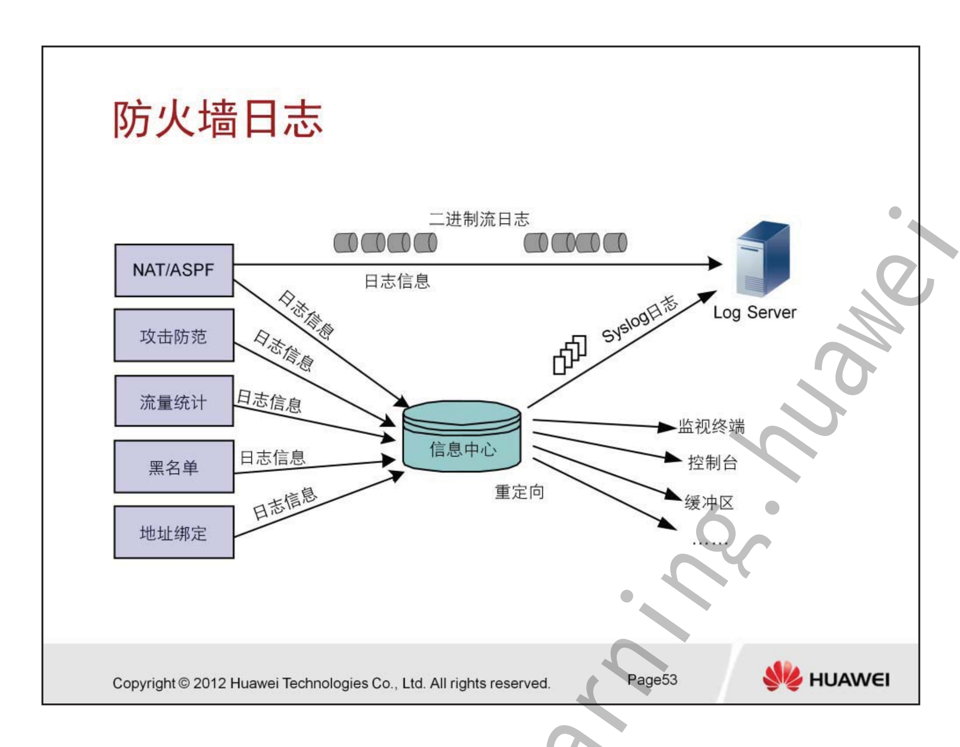
防火墙可以和多个IDS Server联动，Server的IP地址以client address0, client address1等顺序列出。但其它的一些参数如端口，验证参数等都得一致。



目 录

Page52





USG防火墙能够输出以下几种日志信息：

NAT日志和ASPF流日志

日志内容中包含一个完整流的源地址、源端口、目的地址、目的端口等信息，及流的开始和结束时间、流的状态信息等。对于使用NAT功能的流还标识了地址转换之后的地址和端口信息。

攻击防范日志

当发生大量攻击时，USG防火墙利用队列机制对防火墙支持的攻击防范特性提供日志告警信息，通过SYSLOG方式输出告警，告警信息包括攻击来源（源地址）和攻击种类等。

流量监控日志

USG防火墙根据安全域、IP地址等参数进行流量监控，判断速率或连接数目是否达到上限或下限值，当达到上限时触发告警并记录日志，从而有效监控流量；当达到下限时，也触发告警，指示系统恢复正常。

黑名单日志

USG防火墙对于在检测中发现的非法用户，自动将该用户的源IP地址加入到黑名单中，并产生一条黑名单日志，该日志记录主机地址、加入原因等信息。

多种统计信息

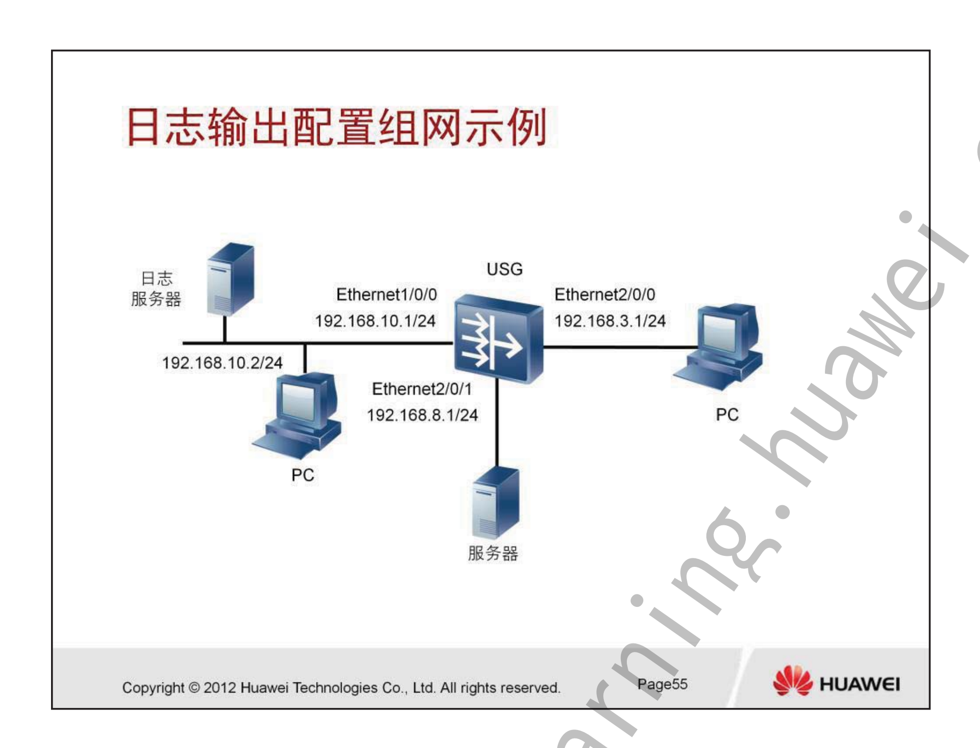
记录流统计信息，了解防火墙运行状况。这些流统计信息包括：总的连接数目、当前连接及半连接数目、最高峰值及丢弃报文数目。记录各种攻击报文的数目，了解攻击事件的发生情况。

地址绑定日志

记录用户配置操作的IP与Mac地址绑定的信息。

在USG防火墙中，攻击防范、流量监控、黑名单和地址绑定产生的日志信息量小，因此采用SysLog方式以文本格式进行输出。这些日志信息必须通过VRP平台的信息中心进行日志管理和输出重定向，或者显示在终端屏幕上，或将SysLog日志发送给日志采集服务器进行存储和分析。

相反，NAT/ASPF产生的日志信息量很大，因此对于这种类型的流提供了一种“二进制”输出方式，直接输出到日志采集服务器上以便对日志进行存储和分析，无需USG上的信息中心模块的参与。相比较而言，二进制流日志的传输效率高于Syslog。



防火墙USG以太网口Ethernet1/0/0配置为Trust域，以太网口Ethernet2/0/0配置为Untrust域，以太网口Ethernet2/0/1配置为DMZ区。防火墙向日志主机输出攻击防范日志信息。

日志输出配置

[USG2100] info-center enable

[USG2100] info-center loghost 192.168.10.2 language english

[USG2100] firewall session log-type binary host 1 192.168.10.2 9002

 Page56



开启信息中心。

[USG2100] info-center enable

配置日志主机IP地址为192.168.10.1。

[USG2100] info-center loghost 192.168.10.2 language english

通过设置日志主机的IP地址，可使信息在该方向输出。语言可选择中文或英文，默认为英文。

对于Nat/ASPF配置二进制流日志输出格式，并配置日志主机的IP地址和接收日志端口。

[USG2100] firewall session log-type binary host 1 192.168.10.2 9002

传输协议为UDP协议。



该命令详细信息显示如下:

[USG2100]display info-center

Information Center:enabled

Log host:

192.168.10.2, channel number 2, channel name loghost,

language english , host facility local7

…

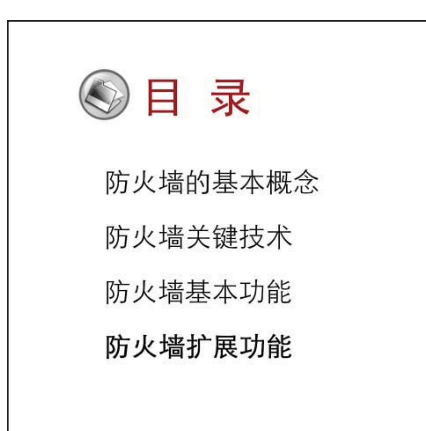
Log buffer:

enabled,max buffer size 1024, current buffer size 512,

current messages 23, channel number : 4, channel name : logbuffer

dropped messages 0, overwritten messages 0

…



Page58




目 录
4. 防火墙扩展功能
4.1 负载均衡
4.2 虚拟防火墙
Copyright © 2012 Huawei Technologies Co., Ltd. All rights reserved.
Page59
HUAWEI

负载均衡

当前的网络应用中，单台服务器的处理能力已经成为网络中的瓶颈，尤其是在IDC、网站等应用场合。

USG防火墙的负载均衡即是将用户流量分配到多个服务器上，从而达到流量分担的目的，进而保障服务器的可用性。防火墙按照配置的算法，将用户流量分配到不同的服务器上，充分利用各个服务器的处理能力，达到最佳的可扩展性。

 Page60

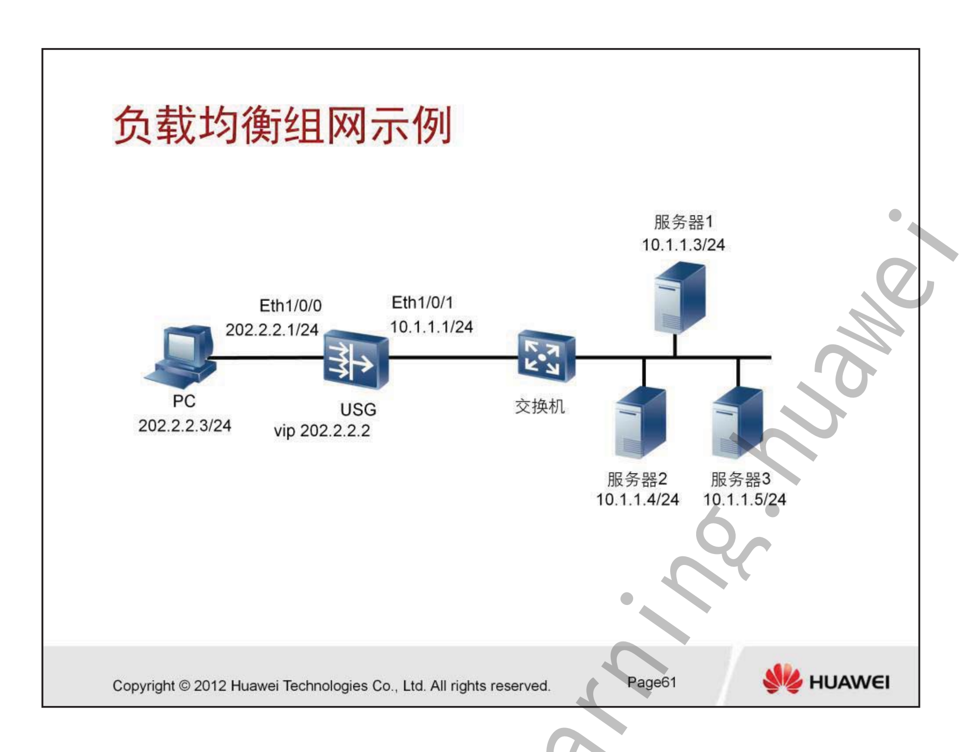


当前的网络应用中，单台服务器的处理能力已经成为网络中的瓶颈，尤其是在IDC、网站等应用场合。具体体现在：单路服务器的平均处理能力仅为1K TPS，而访问服务器的用户却很多。如果单纯升级服务器的性能，则浪费了前期的投资，且费用昂贵；如果增加服务器的数目，会造成控制复杂，容错和热备冗余能力有限，且抗网络DoS攻击能力弱。

USG防火墙的负载均衡即是将用户流量分配到多个服务器上，从而达到流量分担的目的，进而保障服务器的可用性。防火墙按照配置的算法，将用户流量分配到不同的服务器上，充分利用各个服务器的处理能力，达到最佳的可扩展性。

防火墙负载均衡功能解决了单台服务器处理能力有限的问题。所用的负载算法主要是：源地址HASH算法、轮询算法以及加权轮询算法。

典型的应用会将防火墙放在私网出口上。除提供负载均衡功能外，防火墙还能支持常规攻击防范，保障服务器的安全。



内部网络中DMZ区域存在三台真实服务器对外提供FTP服务，IP地址分别为10.1.1.3/24、10.1.1.4/24和10.1.1.5/24。对外的虚拟IP地址为202.2.2.2 。PC位于外网Untrust区域，配置防火墙SLB功能，保证服务器的负载均衡。



rserver rserver-id [to end-rserver-id] rip ip-address [[active | inactive | healthchk] | description text | weight weight]

参数描述：

rserver-id：真实服务器ID，整数形式，范围为0～128。

end-rserver-id：批量配置服务器时指定的最后一个真实服务器ID，整数形式，范围为0～128。注意end-rserver-id要大于rserver-id，否则不能配置。

rip ip-address：设置真实服务器的IP地址。批量配置时，IP地址自动往后加1。如配置ID为1～3的真实服务器，设置rip ip-address为10.100.1.1，则3个真实服务器IP地址分别为10.100.1.1、10.100.1.2和10.100.1.3。

active | inactive：激活/去激活真实服务器。当真实服务器被强制设置激活或去激活状态后，防火墙不对其进行健康性检查。缺省状态为不配置真实服务器的状态，即防火墙对真实服务器进行健康性检查。

healthchk：配置防火墙对真实服务器的健康行检查。

weight weight：设置真实服务器的权重，防火墙可根据服务器指定的权重判断数据流应该流向哪一台服务器。其中，weight为实服务器的权重。

整数形式，范围为1～63。缺省值为32。

description text：设置描述文本。其中，text为服务器描述文本，字符串形式，长度为1～31。

描述：使用rserver配置真实服务器的IP地址。

负载均衡配置 (续)

```
[USG2100 -slb] group group1
[USG2100 -slb-group- group1] metric roundrobin
[USG2100 -slb-group- group1] addrserver 1
[USG2100 -slb-group- group1] addrserver 2
[USG2100 -slb-group- group1] addrserver 3
[USG2100 -slb] vserver huawei vip 202.2.2.2 group group1
```

 Page64



metric { srchash | roundrobin | weightrr }

参数：

- srchash：源地址哈希算法。
- roundrobin：轮询算法。
- weightrr：加权轮询算法。

缺省情况下，采用轮询算法。

vserver server-name vip ip-address group group-name [tcp | udp] [vport port-number [rport port-number]]

参数：

- vserver server-name：虚服务器名称。字符串形式，长度为1～31。
- vip ip-address：虚服务器IP地址。
- group group-name：服务组名称。字符串形式，长度为1～31。
- tcp：使用TCP协议。
- udp：使用UDP协议。

- vport port-number：虚端口号。范围为1～65535。
- rport port-number：实端口号。范围为1～65535。

描述：vserver命令用于配置虚服务器，包括虚服务器的名称和IP地址、对应的服务器组、使用的协议以及协议使用的虚端口号和实端口号。



Slb配置验证：

防火墙上配置了一个虚拟IP202.2.2.2对外提供服务，对应的内部服务器的服务器组名为group1，采用轮询调度机制，里头包含的内部服务器为1，2和3。内部服务器的真实的IP地址分别为10.1.1.3，10.1.1.4和10.1.1.5，并且防火墙会自动对服务器进行健康检查，用Ping检查服务器是否在线。

负载均衡效果

```
[USG2100] display firewall session table
  icmp, (vpn: public -> public) 10.1.1.1:2048-->10.1.1.3:43
  icmp, (vpn: public -> public) 10.1.1.1:2048-->10.1.1.4:43
  icmp, (vpn: public -> public) 10.1.1.1:2048-->10.1.1.5:43
  FTP, (vpn: public -> public) 202.2.2.2:21[10.1.1.3:21]<-+202.2.2.3:1227
  FTP, (vpn: public -> public) 202.2.2.2:21[10.1.1.4:21]<-+202.2.2.3:1229
  FTP, (vpn: public -> public) 202.2.2.2:21[10.1.1.5:21]<-+202.2.2.3:1231
  FTP, (vpn: public -> public) 202.2.2.2:21[10.1.1.3:21]<-+202.2.2.3:1233
  FTP, (vpn: public -> public) 202.2.2.2:21[10.1.1.4:21]<-+202.2.2.3:1235
Current Total Sessions : 8
```

 Page67



从防火墙的Session表项中可以看到，前面三个ICMP表项为防火墙对内部服务器的健康检查，用ping来检测。

后续的表项为从PC FTP提供服务的虚拟IP 202.2.2.2，系统轮询调度连接请求给不同的内部服务器。


目 录
4. 防火墙扩展功能
4.1 负载均衡
4.2 虚拟防火墙
Copyright © 2012 Huawei Technologies Co., Ltd. All rights reserved.
Page68
HUAWEI

虚拟防火墙

在USG上创建逻辑上的虚拟防火墙（Virtual-firewall, Vfw），能够提供防火墙的出租业务，实现子网隔离和解决地址重叠的问题。

每个虚拟防火墙都是VPN实例（VPN-Instance）、安全实例和配置实例的综合体，能够为虚拟防火墙用户提供私有的路由转发平面、安全服务和配置管理平面。

Rfw
Vfw1
Vfw2
Vfw3

 Page69 HUAWEI

随着近年来VPN技术的兴起和不断发展，虚拟防火墙技术应运而生。通过在Eudemon上创建逻辑上的虚拟防火墙（Virtual-firewall,Vfw），能够提供防火墙的出租业务。每个虚拟防火墙都是VPN实例（VPN-Instance）、安全实例和配置实例的综合体，能够为虚拟防火墙用户提供私有的路由转发平面、安全服务和配置管理平面。

VPN实例为虚拟防火墙提供相互隔离的VPN路由，与虚拟防火墙相关的信息主要包括：VPN路由以及与VPN实例绑定的接口。VPN路由将为转发来自与VPN实例绑定的接口的报文提供路由支持。VPN实例与虚拟防火墙是一一对应的。

安全实例为虚拟防火墙提供相互隔离的安全服务，同样与虚拟防火墙一一对应。安全实例具备私有的区域、域间、ACL规则组和NAT地址池，并且能够将绑定接口加入私有区域；安全实例能够为虚拟防火墙提供地址绑定、黑名单、地址转换、包过滤、统计、攻击防范、ASPF和NAT ALG等私有的安全服务。

配置实例为虚拟防火墙用户提供相互隔离的配置管理平面，与虚拟防火墙是一一对应的。虚拟防火墙用户登陆防火墙后有权管理和维护私有的VPN路由和安全实例。

创建虚拟防火墙后，逻辑上，防火墙将分为两类：根防火墙（Root-firewall,Rfw）和虚拟防火墙。根防火墙对应于未配置虚拟防火墙功能的防火墙，并和所有虚拟防火墙构成超级防火墙（Super-firewall,Sfw）。超级防火墙用户有权管理根防火墙和所有虚拟防火墙。

虚拟防火墙

USG 防火墙支持虚拟防火墙特性

- 每个虚拟防火墙均可以独立支持Local、TRUST、UNTRUST、DMZ、VZONE 5个安全区域，接口灵活划分和分配。
- 系统资源独立分配，提供独立的安全业务、NAT多实例、VPN多实例特性。

一台Eudemon物理防火墙
根防火墙 Root FW
虚拟防火墙 Virtual FW
USG
VZONE
Trust
DMZ
VPN-1
UnTrust
Trust
DMZ
VPN-100
UnTrust

Page71
HUAWEI

USG系列防火墙支持虚拟防火墙功能，不同的设备型号，支持的数量指标不一样，有的为100，有的为512等。

在VPN网络环境的时候，可以使用一台防火墙设备为多个VPN独立的提供安全服务。USG系列防火墙的VPN服务＋强大的安全业务可以为用户构建安全可靠的VPN系统服务。

每个虚系统可以划分多个物理接口或者是VLAN子接口，每个虚系统可以独立拥有五个安全区域：Local, trust、untrust、DMZ、VZONE。其中VZONE安全区域定义了从一个虚系统到另外虚系统之间的接口，通过VZONE，一个虚系统可以方便的控制经过本虚系统到外部虚系统之间的流量。Eudemon防火墙的虚拟防火墙系统可以分为“根系统”和“虚拟防火墙”两种类型。“根防火墙”在虚拟防火墙系统中主要是连接公网系统（Internet），“虚拟防火墙”连接各个VPN系统。

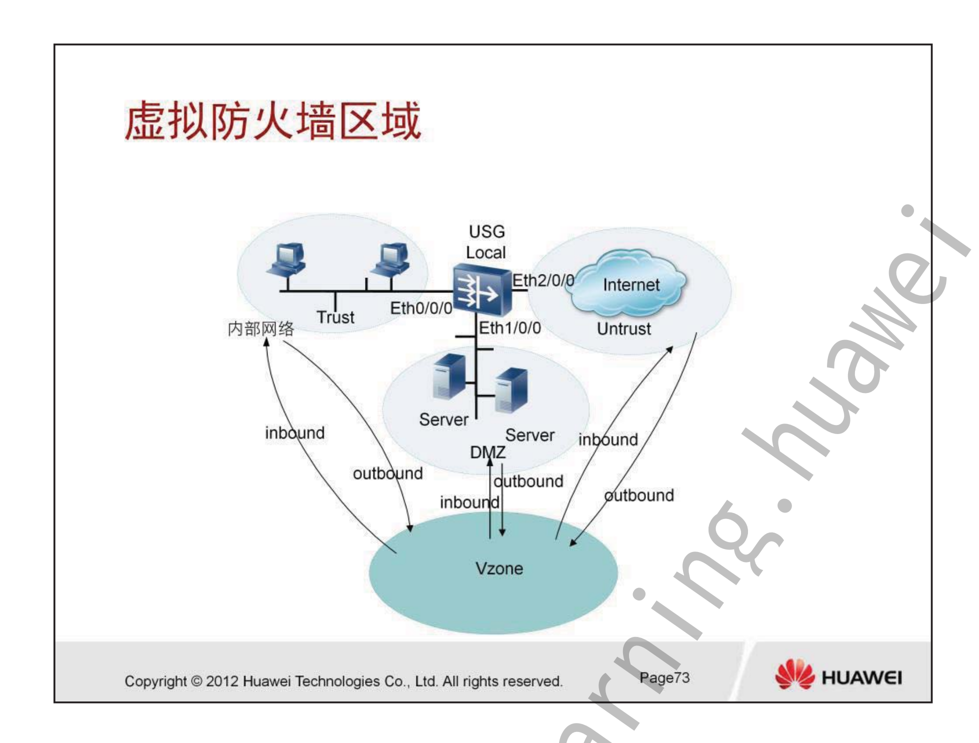
USG的虚拟防火墙的特点：

资源独立分配，每个虚系统的转发表项等资源是独立分配的，这样从技术上就保证了每一个虚系统和一个独立防火墙从实现上是一样的。而且非常安全，各个虚系统之间是无法直接访问的。只有在虚系统之间引入了对方的某些路由，才可以使得虚系统之间是可以通信的。

引入了VZONE（虚拟安全区域）的概念，VZONE是虚系统之间访问的策略配置的必要关口，每个虚系统都有一个VZONE。需要互通的两个虚系统之间只有都打开VZONE对应的安全策略，才可能保证两个安全区域之间的访问。这样就大大增加了虚系统之间的安全。

所有的安全业务都可以针对虚系统独立配置。例如，包过滤、NAT、攻击防范等等。这样在用户使用虚系统服务的时候，也可以同时享受各种USG所能提供的各种安全业务。

每个虚系统提供独立的管理员权限，针对虚系统的管理员只能管理各自虚系统的相关配置，也只能看到自己虚系统的配置。对虚系统管理员而言，他只能看见一个虚拟独立的防火墙配置。

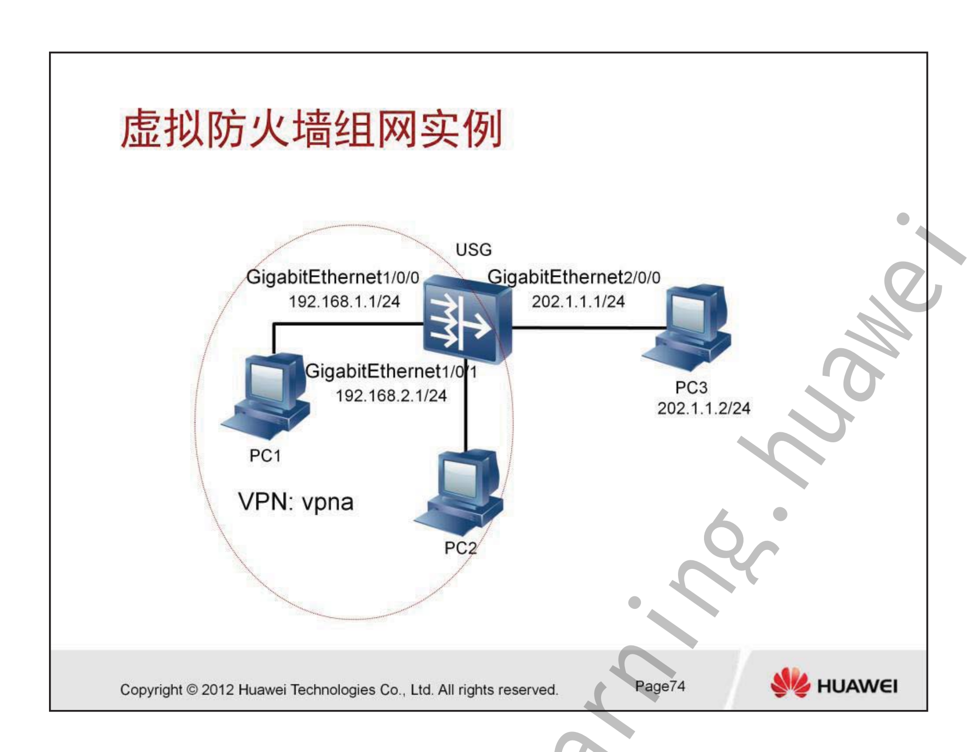


虚拟防火墙与根防火墙相同，系统缺省保留五个安全区域，分别为本地区域（Local）、受信区（Trust）、非军事化区（DMZ）、非受信区（Untrust）和虚拟区域（Vzone）。

一个VPN instance缺省包含local（安全优先级100） trust（85） untrust（5） dmz（50） vzone（0） 五个安全区域，每个vpn instance都包含一个vzone，但，vzone与local是不同的。每个vpn instance都有一个对应的vzone，从理论上说，vzone间是互相连通的，因此在配置跨vpn转发的时候，只需要在一个vpn实例中配置trust-vzone的包过滤关系，再在另一个vpn实例中配置vzone-untrust的包过滤关系，就可以使会话跨越VPN实例了。

在防火墙中，当报文从高优先级区域向低优先级区域发起连接时，如从根防火墙的Trust区域向Untrust区域发起数据连接以及虚拟防火墙中，从DMZ区向Untrust区发起数据连接时，必须明确配置缺省过滤规则。

由防火墙本地（Local区域）发起或终止的报文不进行状态检测，这类报文的过滤由包过滤机制来完成。



PC1 和 PC2属于 vpna，PC3属于公网，要求vpna的用户经过地址转换能够访问公网中的地址PC3。

接口IP等的信息如图所示。

虚拟防火墙配置—接口

```
[USG2100]ip vpn-instance vpna
[USG2100 -vpn-vpna]route-distinguisher 100:1
[USG2100]int GigabitEthernet1/0/0
[USG2100-GigabitEthernet1/0/0]ip binding vpn-instance vpna
[USG2100-GigabitEthernet1/0/0] ip address 192.168.1.1 255.255.255.0
[USG2100]int GigabitEthernet 1/0/1
[USG2100 -GigabitEthernet1/0/1]ip binding vpn-instance vpna
[USG2100 -GigabitEthernet1/0/1] ip address 192.168.2.1 255.255.255.0
…
```


Page75



虚拟防火墙的配置和普通的防火墙配置基本一致，有一点需要注意的是凡是加入到虚拟安全区域的接口都必须先绑定到该虚拟实例，然后再配置相应的IP地址。

省略号的主要操作为把接口添加进相应的区域。

虚拟防火墙配置－安全策略

```
[USG2100] acl number 2000  vpn-instance vpna
[USG2100 -acl-basic-2000]rule permit
[USG2100]firewall interzone vpn-instance vpna trust untrust
[USG2100 -interzone-trust-untrust-vpna]packet-filter 2000
  inbound
[USG2100 -interzone-trust-untrust-vpna]packet-filter 2000
  outbound
```

 Page76 

对于在同一个虚拟防火墙内部的各区域之间的通信控制，和普通的防火墙安全策略配置一致。

被VPN实例引用的ACL必须关联到相应的VPN实例。

跨VPN实例访问配置

```
[USG2100] nat address-group 1  202.1.1.3 202.1.1.8 vpn-instance
   vpna
[USG2100] firewall interzone vpn-instance vpna trust vzone
[USG2100 -interzone-trust-vzone-vpna] packet-filter 2000 outbound
[USG2100 -interzone-trust-vzone-vpna] nat outbound 2000 address-
   group 1
[USG2100]ip route-static vpn-instance vpna 0.0.0.0 0 202.1.1.2
   public
```

 Page77



如果需要跨VPN实例访问，比如vpna的trust区域需要访问根防火墙的untrust外网。数据流的首包流向是：vpna的trust区域->Vzone区域->根防火墙的untrust区域，期间在vpna的trust至vzone方向做一次地址转换，为此我们还需要为vpna的路由表添加一条路由，下一跳为公网的出口。至于返回的报文，根据绑定vpn实例的地址池中的地址，自动对应上vpn实例，还原回相应的地址映射关系，并查vpn的IP路由表转发。

备注：其它省略的命令主要为根防火墙untrust区域的接口配置及vzone->untrust方向的安全策略，为实验方便，我们简单处理，默认报文全部允许通过。



参考:

- 防火墙是状态防火墙，维护连接信息，路由器不维护连接信息；
- 面向的功能不一样，一个是通信控制，一个是互联互通；
- 防火墙在针对数据的安全性方面作了更多的检查过滤功能；
- 防火墙有安全域的概念，而路由器只有接口的概念。

谢谢

www.huawei.com


USG防火墙NAT业务
特性与配置
www.huawei.com
Copyright © 2012 Huawei Technologies Co., Ltd. All rights reserved.
HUAWEI



前 言

随着因特网的快速发展，它所面临的两个最迫切的问题就是IP地址的匮乏和路由规模的扩大。对此，长期的和短期的解决方案都有所发展，那就是网络地址转换（NAT）和IPv6（下一代因特网协议）技术。在IPv6技术尚在研究中且还未完全取代现有的IPv4网络的情况下，短期解决方案——NAT技术对于缓解目前的地址缺乏问题显得尤为重要。

 Page1





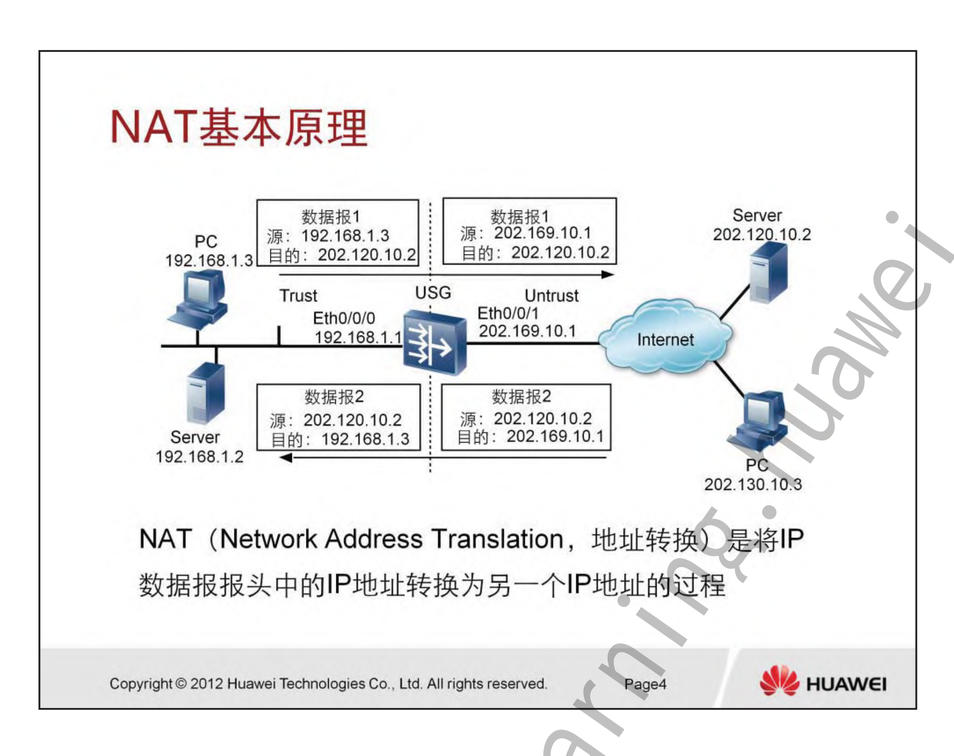
培训目标

学完本课程后，您应该能：

- 描述NAT技术的原理
- 描述USG防火墙的各种NAT特性
- 描述USG防火墙各NAT特性的基本组网与配置

Page2
HUAWEI

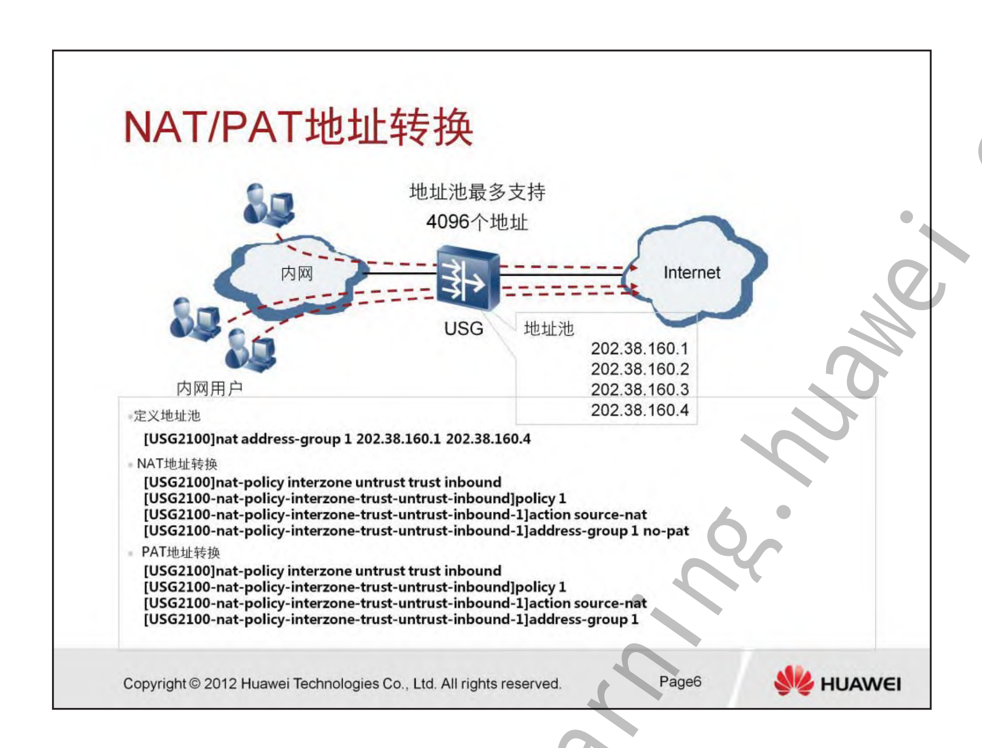

目 录
NAT技术原理
USG防火墙NAT特性与配置
Copyright © 2012 Huawei Technologies Co., Ltd. All rights reserved.
Page3
HUAWEI



地址转换分为两种方式：

- 1：1转换
- 私网地址和公网地址一一对应，不转换端口。
  - 优点：实现简单；
  - 缺点：一个私网地址需要一个公网地址与之对应，并不能解决公网地址短缺问题。
- N：1转换
- 即同时转换地址和端口，支持多个私网地址映射为同一个公网地址，可以解决公网地址短缺问题。


目 录
NAT技术原理
USG防火墙NAT特性与配置
Copyright © 2012 Huawei Technologies Co., Ltd. All rights reserved.
Page5
HUAWEI

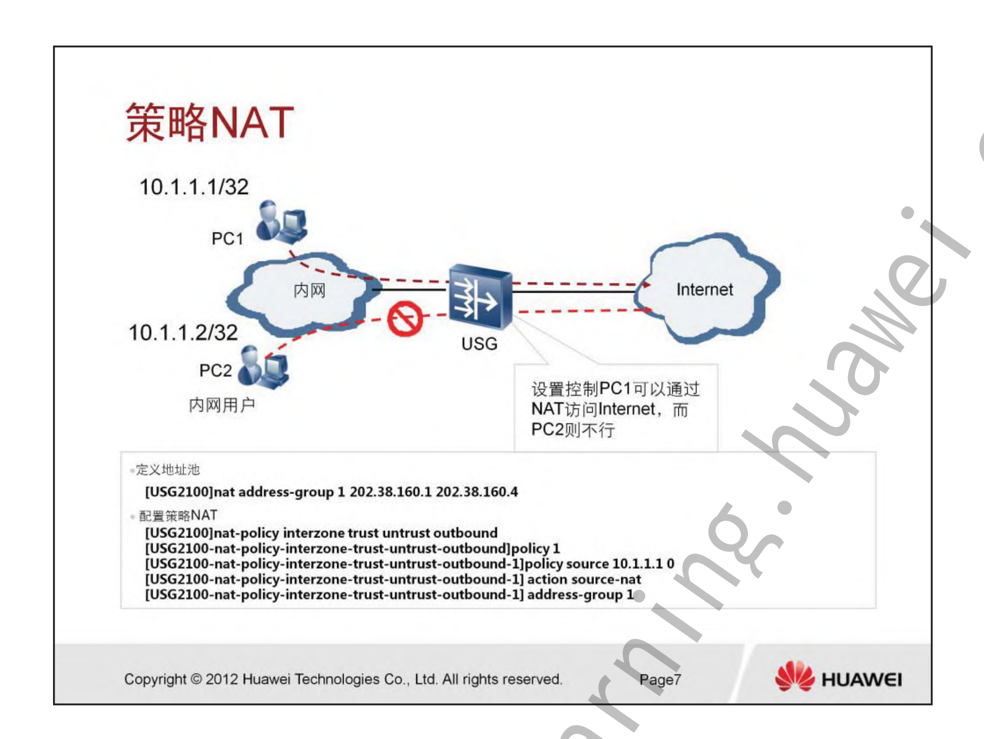


地址转换支持地址池的概念，在转换时，从地址池中根据一定的规则选择一个IP地址做为转换后的源地址，完成地址转换。这个选择的过程对用户来讲是透明的，用户只需要将他具有的IP地址配置成相应的地址池就可以了。对于使用地址池中的地址进行转换可以有三种形式：

- NAT：1对1的地址翻译，静态NAT，内网中的每个主机都被永久映射成外网中的某个合法的地址，多用于服务器
- PAT：多个内网地址翻译到1个外网IP地址，把内部地址映射到外网的一个IP地址的不同端口上（有些协议必须保持端口不变，如NETBIOS。在NAT转换时，防火墙会自动识别这些协议。）
- NPAT：多个内部网地址翻译到N个IP地址池中的地址

USG2200/5100/5500、USG2200/5100 BSR/HSR最多支持1024个地址池，每个地址池最多可配置4096个公网地址；USG2110-X/2100、USG2100 BSR/HSR最多支持24个地址池每个地址池最多可配置4096个公网地址。

专利技术：在转换时不单单记录报文的源地址/端口转换信息，防火墙还会记录报文的目的IP/端口信息，这样就可以用相同的IP/端口提供对不同目的IP/端口的报文的NAT转换了。



USG防火墙的地址转换功能，可以利用访问控制列表决定什么样的地址可以进行地址转换。如果某些主机具有访问Internet的权利，而某些主机不能访问Internet。可以利用policy定义什么样的主机不能访问Internet，什么样的主机可以访问Internet。然后将配置好的policy规则应用在地址转换上，就可以达到利用policy控制地址转换的功能。

如图，PC2不需要访问Internet，那么可以通过policy进行控制，限制PC2访问Internet。使得PC2只能访问内部局域网的主机。

NAT ALG功能

NAT和NAPT只能对IP报文的头部地址和TCP/UDP头部的端口信息进行转换。对于一些特殊协议，例如ICMP、FTP等，它们报文的数据部分可能包含IP地址或端口信息，这些内容不能被NAT有效的转换，就可能导致问题。

例如，一个使用内部IP地址的FTP服务器可能在和外部网络主机建立会话的过程中需要将自己的IP地址发送给对方。而这个地址信息是放到IP报文的数据部分，NAT无法对它进行转换。当外部网络主机接收了这个私有地址并使用它，这时FTP服务器将表现为不可达。

 Page8

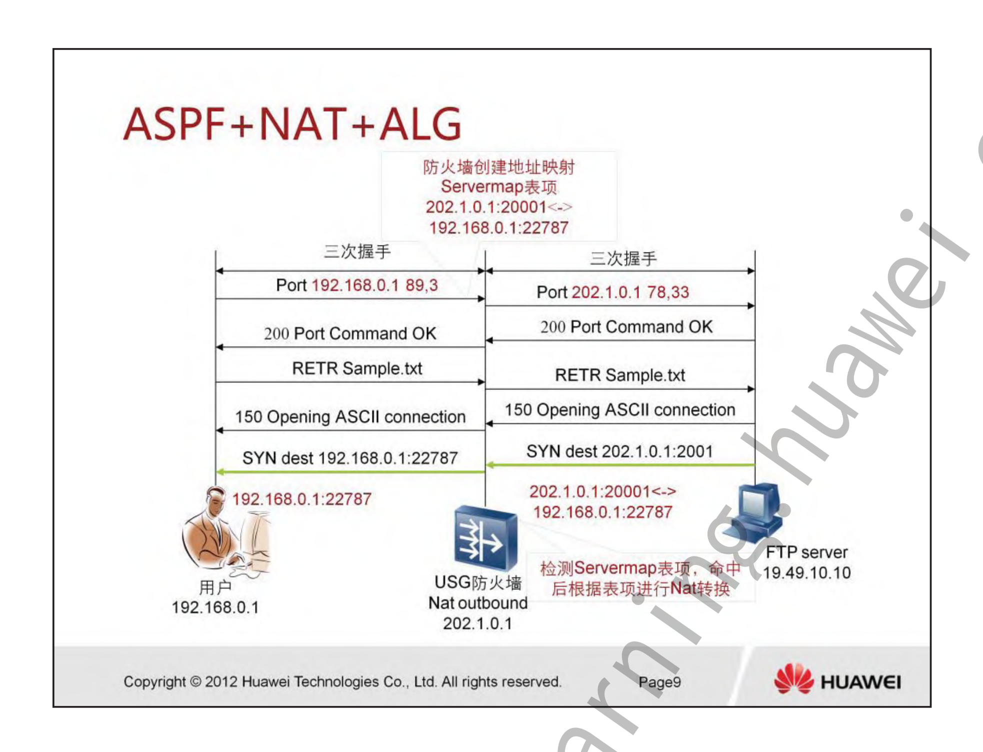


解决这些特殊协议的NAT转换问题的方法就是在NAT实现中使用ALG（Application Level Gateway）功能。ALG是特定的应用协议的转换代理，它和NAT交互以建立状态，使用NAT的状态信息来改变封装在IP报文数据部分中的特定数据，并完成其他必需的工作以使应用协议可以跨越不同范围运行。

USG提供了完善的地址转换应用级网关机制，使其在流程上可以支持各种特殊的应用协议，而不需要对NAT平台进行任何的修改，具有良好的可扩充性。目前它所实现了常用应用协议的ALG功能包括：

- DNS；FTP；H.323；HWCC（Huawei Conference control Protocol）；ICMP；ILS（Internet Locator Service）；MGCP（Media Gateway Control Protocol）；MSN；NetBIOS；PPTP；QQ；RTSP（Real Time Streaming Protocol）等。

NAT ALG功能的使能，在域间使用 detect Protocol，就会使能对应协议的NAT ALG功能。



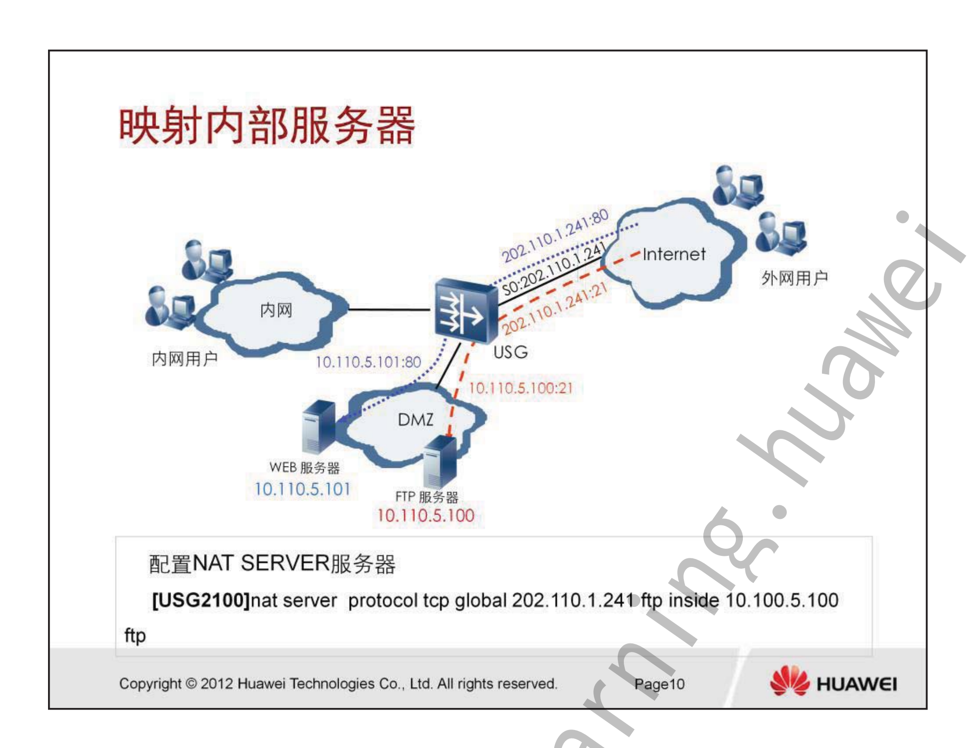
FTP包含一个预知端口的TCP控制通道和一个动态协商的TCP数据通道，在进行Nat操作的时候，Nat设备需要监控并替换控制通道信息，并根据替换的对应关系创建用于NAT转换的Server Map表项；对于一般的NAT设备来说，如果不对控制报文进行相关的Nat Alg处理将相关控制命令替换成NAT后的控制命令的话，服务器端可能会拒绝该命令，如果不建立NAT转换关系的对应表项的话，后续报文将无法转换也就无法正确的访问client。

NAT ALG相关功能则解决了这一问题，它检测协议控制通道信息，动态地对报文进行内容替换并创建和删除临时的Server map表项，以确保数据通道报文Nat转换的正确性。

防火墙的命令配置的NAT SERVER也会产生Server map表，这个表项是永久的。除非将NAT SERVER配置删掉。ALG和ASPF处理的时候，生成Server map表项，都是临时表项，有老化时间，在老化时间内没有访问就会删除。

一般ALG和ASPF建立Server map表后，动态协商的数据流的首报文命中Server map表项后，防火墙会建立SESSION表项，后续的数据报文直接通过SESSION表项转发，SERVER MAP表项就不再被命中，慢慢的老化删除掉。

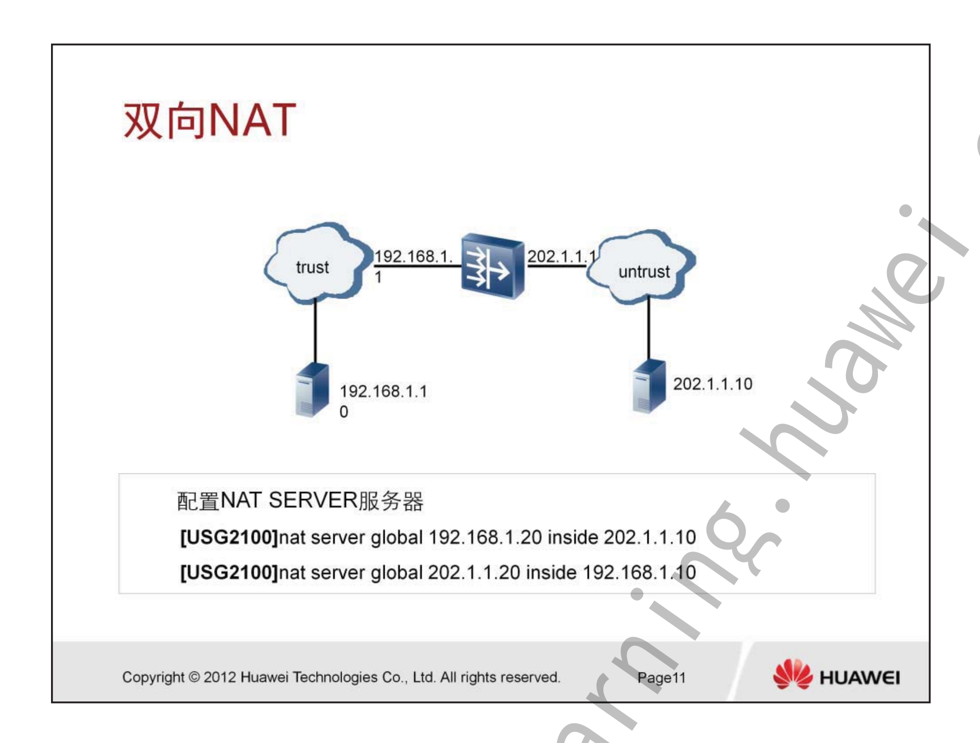
reset firewall session table命令在删除SESSION表的同时，也会将Server map表项删除。



支持私有地址网络的主机可以访问外部网络和外部网络访问内部服务器，是NAT完成的主要功能，也是必须要提供的功能。

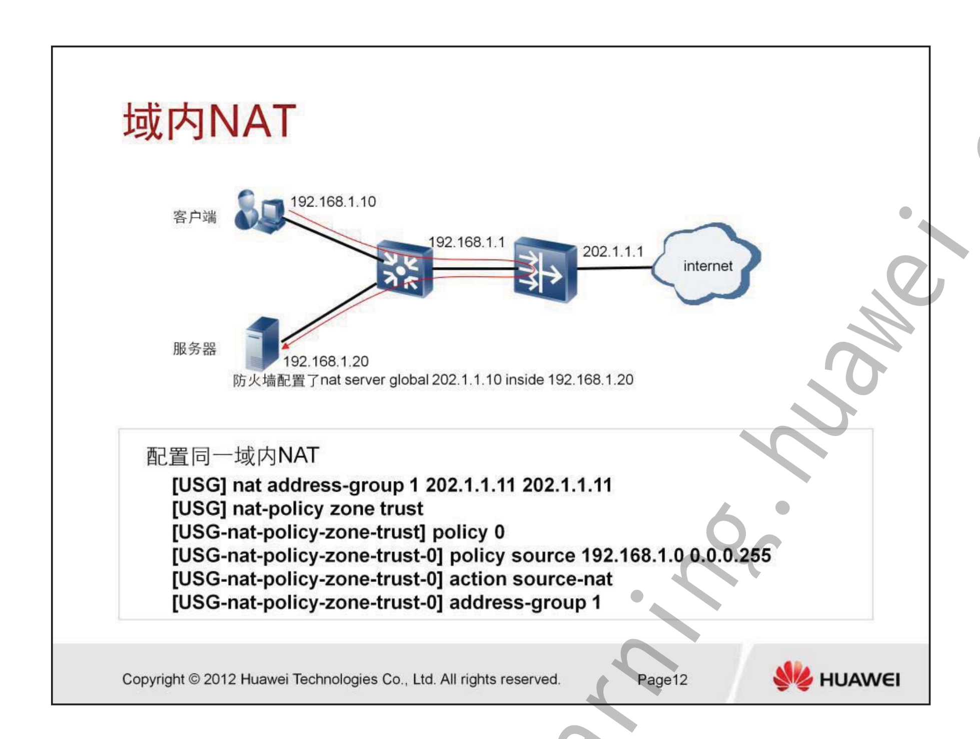
支持NAT Server模式，可以向外映射内部服务器；支持1对多的映射方式，提供端口级的NAT Server模式，可以将服务器的端口映射为外部的一个端口，阻塞服务器的其它端口，增加服务器的安全性。

USG系列防火墙借助NAT提供映射内部服务器功能， 1个公网地址最多支持映射256个服务器私网地址。



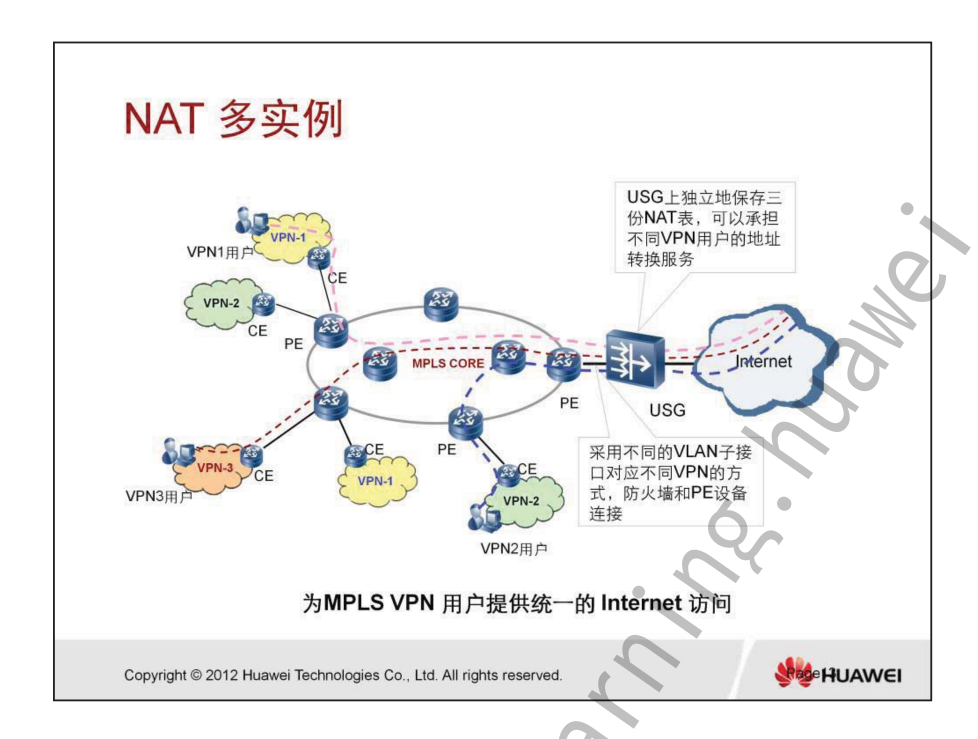
当低优先级区域的用户访问NAT SERVER的公网地址时，会将报文的目的地址转换为内部服务器的私网地址，但内部服务器需要配置到该公网地址的路由。如果要简化配置，避免配置到公网地址的路由，则可以配置从低优先级区域到高优先级区域方向的NAT，即inbound方向的NAT。

同一个安全区域内的访问需要作NAT，则需要配置域内NAT功能。



内网服务器IP地址为私网IP，在防火墙上对其配置了NAT SERVER，对此私网IP地址实现1对1的映射。

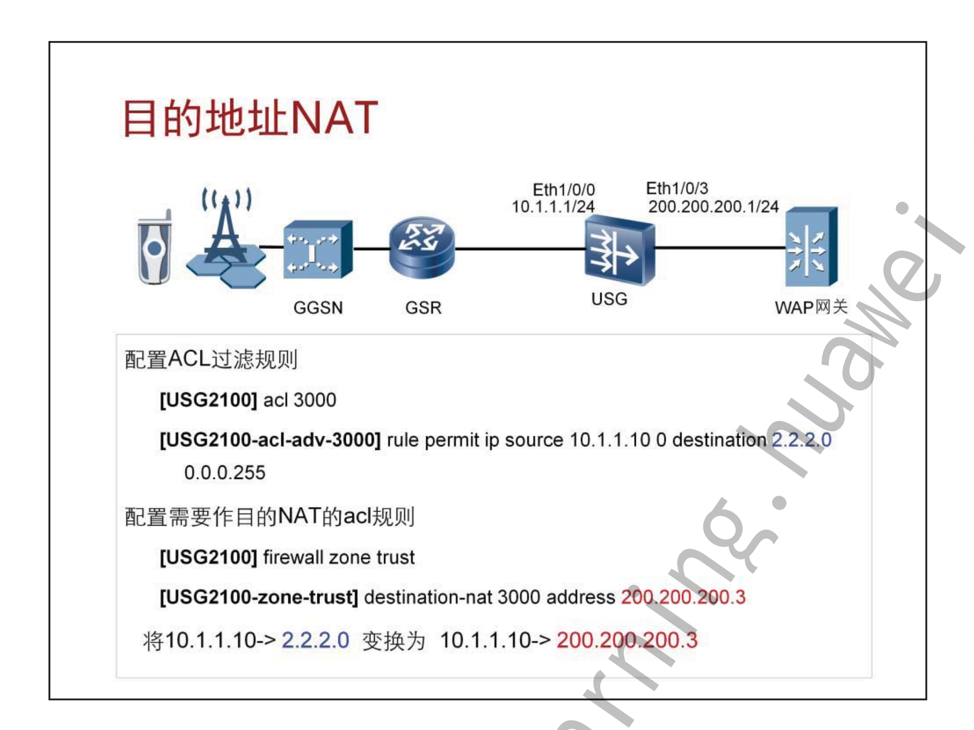
内网客户端访问NAT SERVER的公网地址的时候， 数据包目的服务器IP地址被转换为私网地址，数据报被转发到内网服务器。此时需要在内网的安全域上配置域内NAT，这样才能保证业务正常。



USG系列防火墙支持虚拟防火墙、地址转换多实例、multi-VRF等VPN隔离技术，因此非常适合将USG防火墙放置在一个MPLS骨干网络的出口，做为一个MPLS VPN网络统一访问Internet的出口设备。

USG系列防火墙具有如下一些特点，适合完成如下工作：

- 支持完善的安全防范业务，可以避免来自Internet的攻击，保护整个MPLS网络的安全。
- 支持完善的VLAN、VRF、OSPF、BGP等技术，可以很好的解决MPLS网络和IP网络的融合问题。
- 支持业务多实例特性，资源独立存放，可以很好的解决私网地址重叠的问题。



目前手机上网的需求越来越多，有些手机来自海外，并且它们的WAP网关地址都是写死到手机中的，不能进行人为更改，因此这些手机在国内是不能直接登陆到WAP网关上网的，为了满足这些手机上网的需求，增加了目的NAT特性，当目的地址为这些手机的WAP网关地址的数据报通过防火墙时，防火墙根据匹配的ACL规则将这些地址修改为运营商网络中实际的WAP网关地址，从而使这些手机能够正常登陆WAP网关。

有些来自海外的、手机的WAP网关地址为内网地址，例如10.0.0.1，192.168.1.1等，这些地址不能进行修改，但是一般国内运营商的WAP网关地址不是这些地址，因此需要相关的设备将上述地址转换为实际的WAP网关地址。

注意事项：

由于所有的手机业务都要通过USG，因此要注意区分是WAP业务，还是流媒体或其他业务，一般WAP业务的目的地址用的是私网地址，流媒体等其他业务是公网地址，如果有除了WAP业务之外的其他业务需要配置相应的静态路由，并且在创建需要作目的NAT的规则时要考虑到是否是WAP业务。

基本原理：

手机上网业务主要由移动运营商的WAP网关提供，WAP网关地址在手机中已经配置，手机连接WAP网关时需要经过GGSN，GGSN相当于电信网络连接Internet的一个网关。当手机访问WAP网关时，GGSN会为每一个手机用户分配一个IP地址，数据包通过GSR路由器到达USG防火墙，USG然后将目的地址为WAP网关的数据包转发到WAP网关。如果手机的上网地址不是WAP网关时，需要配置相应的规则，将这些上网请求的目的地址转换为WAP网关的地址。

普通命令：

destination-nat acl-num address X.X.X.X [port port-num]

destination-nat acl-num address X.X.X.X [port port-num]命令用来在域上配置需要作目的NAT的功能。



参考:

- 优点：可以节省已经不够用的IP资源；还有可以隐藏内网的源机器，保护它不容易受到外网的攻击。
- 缺点：保护内网的同时，也使从外网来的访问变得麻烦；另外也使网络追踪变得麻烦；多一层处理，可能多引入一个性能瓶颈。

谢谢

www.huawei.com


USG防火墙攻击防范业务特性与配置
www.huawei.com
Copyright © 2012 Huawei Technologies Co., Ltd. All rights reserved.
HUAWEI



前 言

基于防火墙的组网位置和功能上看，对一些非法攻击的防御是防火墙设备的一个非常重要的功能，通过防火墙的攻击防范的防御功能可以保证内部网络的安全，在这一点上是其他数据通信设备无法替代的，因此在全网解决方案中，防火墙是必不可少的一个部件。

本章主要描述了基于IP的各种网络攻击方式的原理及其在USG防火墙上的防范配置。



防火墙是设置在被保护网络和外部网络之间的一道屏障，以防止发生不可预测的、潜在破坏性的侵入。防火墙由设置在不同网络（如可信任的企业内部网和不可信的公共网）或网络安全域之间的一系列部件的组合构成。它是不同网络或网络安全域之间信息的唯一出入口，能根据企业的安全政策控制（允许、拒绝、监测）出入网络的信息流，且本身具有较强的抗攻击能力。防火墙可通过监测、限制、更改跨越防火墙的数据流，尽可能地对外部屏蔽网络内部的信息、结构和运行状况，以此来实现网络的安全保护。

在提供丰富的攻击防御的手段来保证其他网络设备安全的同时，防火墙本身必须具有非常好的性能表现，否则即使攻击可以勉强发现，但是防火墙本身资源耗尽反而成为了网络的瓶颈，造成了新的故障点。

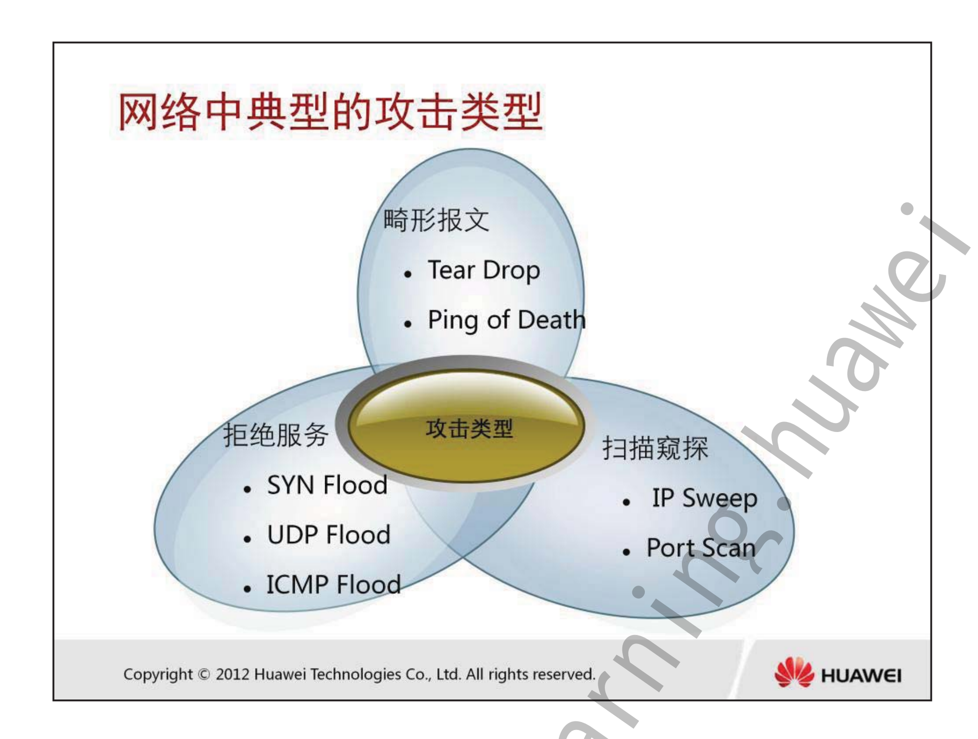


培训目标

学完本课程后，您应该能：

- 描述 IP网络中各种攻击的原理
- 掌握 USG防火墙的各种攻击防范的配置





网络攻击可分为拒绝服务型攻击、扫描窥探攻击和畸形报文攻击三大类
:

拒绝服务型攻击

DoS（Deny of Service）攻击是使用大量的数据包攻击系统，使系统无法接受正常用户的请求，或者主机挂起不能提供正常的工作。主要DoS攻击有SYN Flood、Fraggle等。拒绝服务攻击和其他类型的攻击不同之处在于：攻击者并不是去寻找进入内部网络的入口，而是阻止合法用户访问资源或路由器。

DDoS（Distributed Denial of Service）攻击是一种DoS攻击。这种攻击是使用攻击者控制的几十台或几百台计算机攻击一台主机，使系统无法接受正常用户的请求，或者挂起不能正常的工作。

扫描窥探攻击

扫描窥探攻击是利用ping扫射（包括ICMP和TCP）来标识网络上存活着的系统，从而准确的指出潜在的目标；利用TCP和UDP端口扫描，就能检测出操作系统和监听者的潜在服务。攻击者通过扫描窥探就能大致了解目标系统提供的服务种类和潜在的安全漏洞，为进一步侵入系统做好准备。

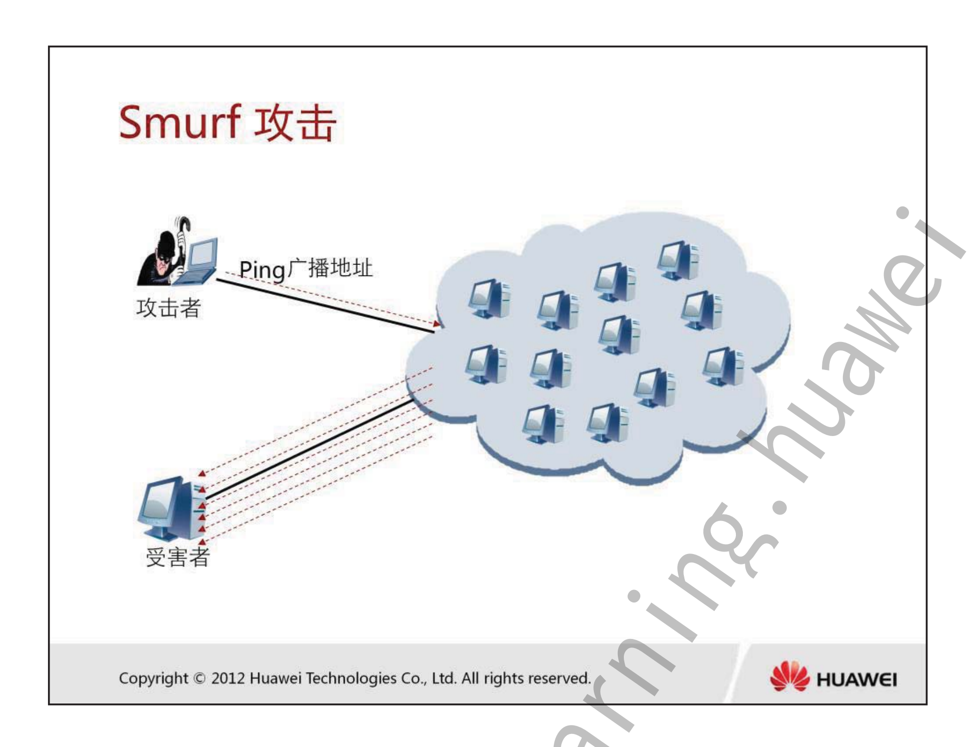
畸形报文攻击

畸形报文攻击是通过向目标系统发送有缺陷的IP报文，使得目标系统在处理这样的IP包时会出现崩溃，给目标系统带来损失。主要的畸形报文攻击有Ping of Death、Teardrop等。



目 录





攻击介绍：

- Smurf攻击方法是发ICMP应答请求，该请求包的目标地址设置为受害网络的广播地址，这样该网络的所有主机都对此ICMP应答请求作出答复，导致网络阻塞。高级的Smurf攻击，主要用来攻击目标主机。方法是将上述ICMP应答请求包的源地址改为受害主机的地址，最终导致受害主机雪崩。攻击报文的发送需要一定的流量和持续时间，才能真正构成攻击。理论上讲，网络的主机越多，攻击的效果越明显。

处理方法：

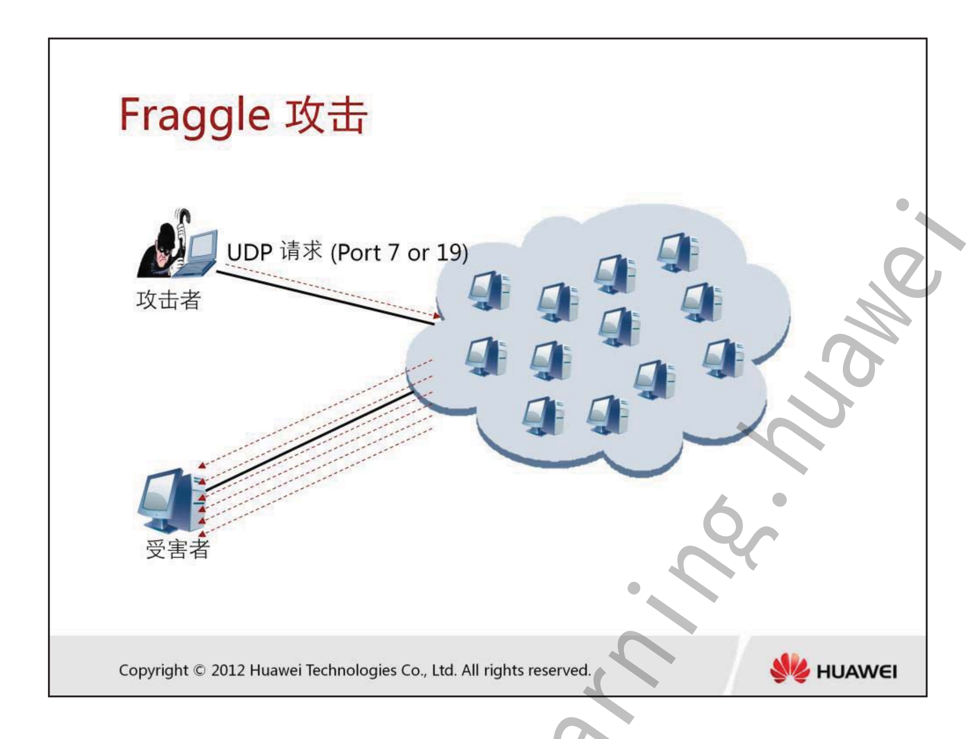
- 检查ICMP应答请求包的目的地址是否为子网广播地址或子网的网络地址，如是，则直接拒绝，并将攻击记录到日志。

配置：

```
firewall defend smurf enable
```

使用限制：

- 由于路由器等三层设备本身就不会转发目的地址是广播地址的报文，因此SMURF攻击在网络上很难形成攻击。在防火墙上，检查SMURF攻击必须要求被攻击网络是直接连接到防火墙上。



攻击介绍：

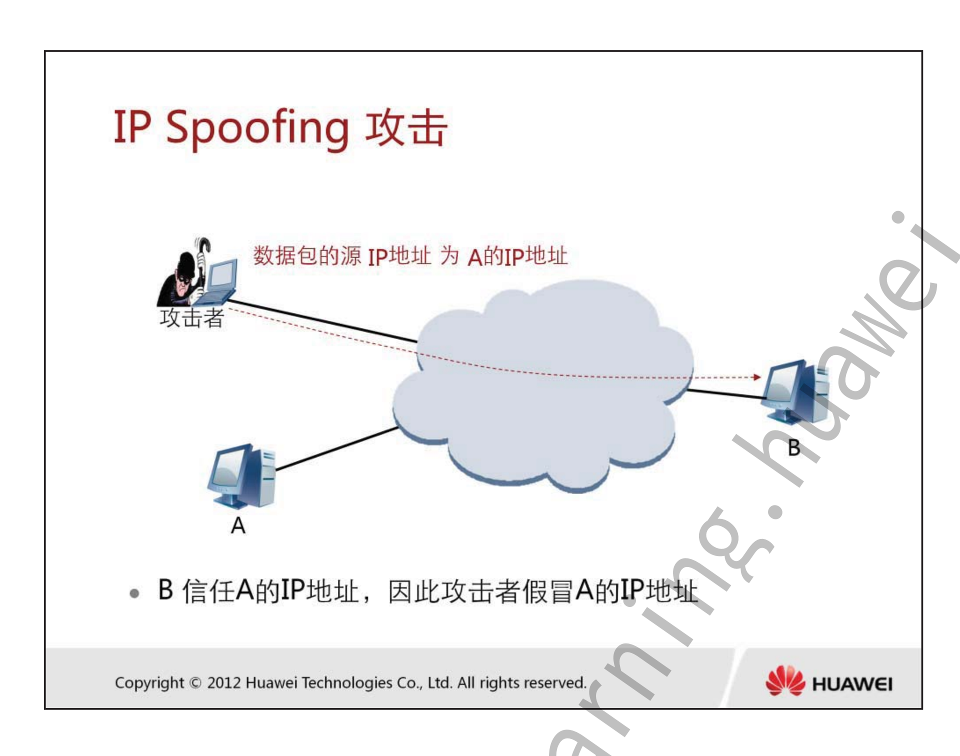
- 类似于Smurf，使用UDP应答消息而非ICMP。UDP端口7（ECHO）和端口19（Chargen）在收到UDP报文后，都会产生回应。在UDP的7号端口收到报文后，会回应收到的内容，而UDP的19号端口在收到报文后，会产生一串字符流。它们都同ICMP一样，会产生大量无用的应答报文，占满网络带宽。 攻击者可以向子网广播地址发送源地址为受害网络或受害主机的UDP包，端口号用7或19。子网络启用了此功能的每个系统都会向受害者的主机作出响应，从而引发大量的包，导致受害网络的阻塞或受害主机的崩溃；子网上没有启动这些功能的系统将产生一个ICMP不可达消息，因而仍然消耗带宽。 也可将源端口改为Chargen，目的端口为ECHO，这样会自动不停地产生回应报文，其危害性更大。

处理方法：

检查进入防火墙的UDP报文，若目的端口号为7或19，则直接拒绝，并将攻击记录到日 志，否则允许通过。

配置：

```
firewall defend fraggle enable
```



攻击介绍:

- 为了获得访问权，入侵者生成一个带有伪造源地址的报文。对于使用基于IP地址验证的应用来说，此攻击方法可以导致未被授权的用户可以访问目的系统，甚至是以root权限来访问。即使响应报文不能达到攻击者，同样也会造成对被攻击对象的破坏。这就造成IP Spoofing攻击。

处理方法:

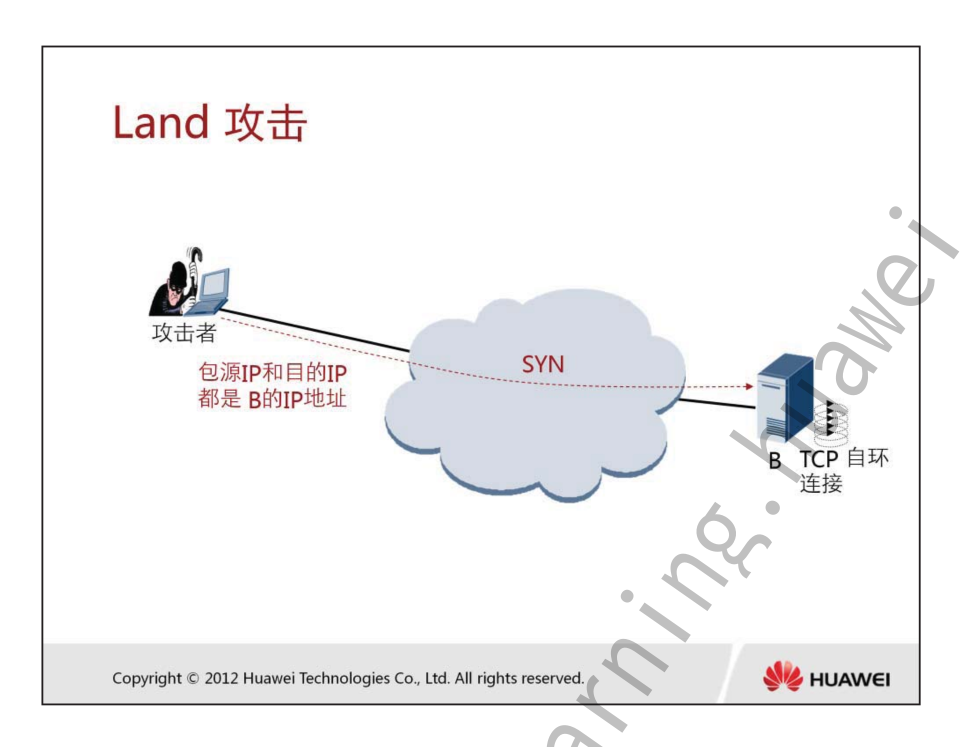
- 检测每个接口流入的IP报文的源地址与目的地址，并对报文的源地址反查路由表，以该报文的源地址作为目的地址查找路由表对应的出接口，如果该出接口与已收到的报文不符，则视为IP Spoofing攻击。将被拒绝，并进行日志记录。

配置:

- firewall defend ip-spoofing enable

使用限制:

- 当报文存在多出口路径的时候或者是存在默认路由的时候，ip spoofing攻击防范效果很差或者是容易出现问题。在实际使用中ip spoofing攻击的用途不大，没有必要打开。



攻击介绍:

- 所谓Land攻击，就是把TCP SYN包的源地址和目标地址都设置成某一个受害者的IP地址。这将导致受害者向它自己的地址发送SYN-ACK消息，结果这个地址又发回ACK消息并创建一个空连接，每一个这样的连接都将保留直到超时掉。各种受害者对Land攻击反应不同，许多UNIX主机将崩溃，NT主机会变的极其缓慢。

处理方法:

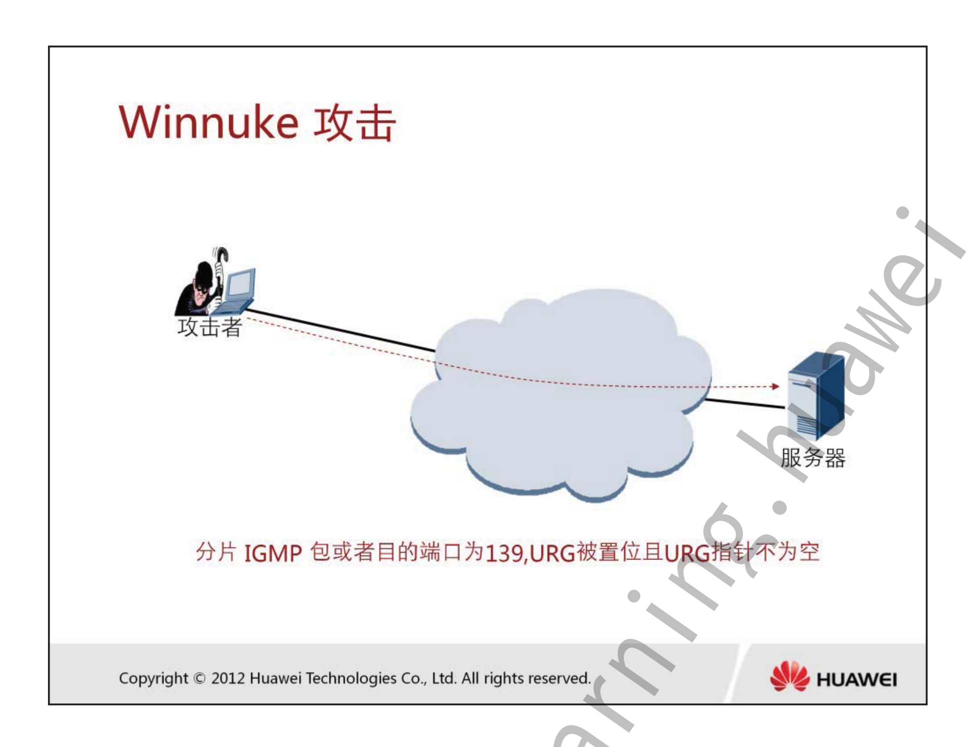
- 对每一个的IP报文进行检测，若其源地址与目的地址相同，或者源地址为环回地址(127.0.0.1)，则直接拒绝，并将攻击记录到日志。

配置:

- firewall defend land enable

使用限制:

- 没有限制，对性能也基本无影响，对网络也不会造成不良影响。



攻击介绍：

- WinNuke攻击又称带外传输攻击，它的特征是攻击目标端口，被攻击的目标端口通常是139、138、137、113、53，URG位设为“1”且URG指针不为空，即紧急模式。还有一种是IGMP分片报文，一般情况下，IGMP报文是不会分片的，所以，不少系统对IGMP分片报文的处理有问题。如果收到IGMP分片报文，则基本可判定受到了攻击。

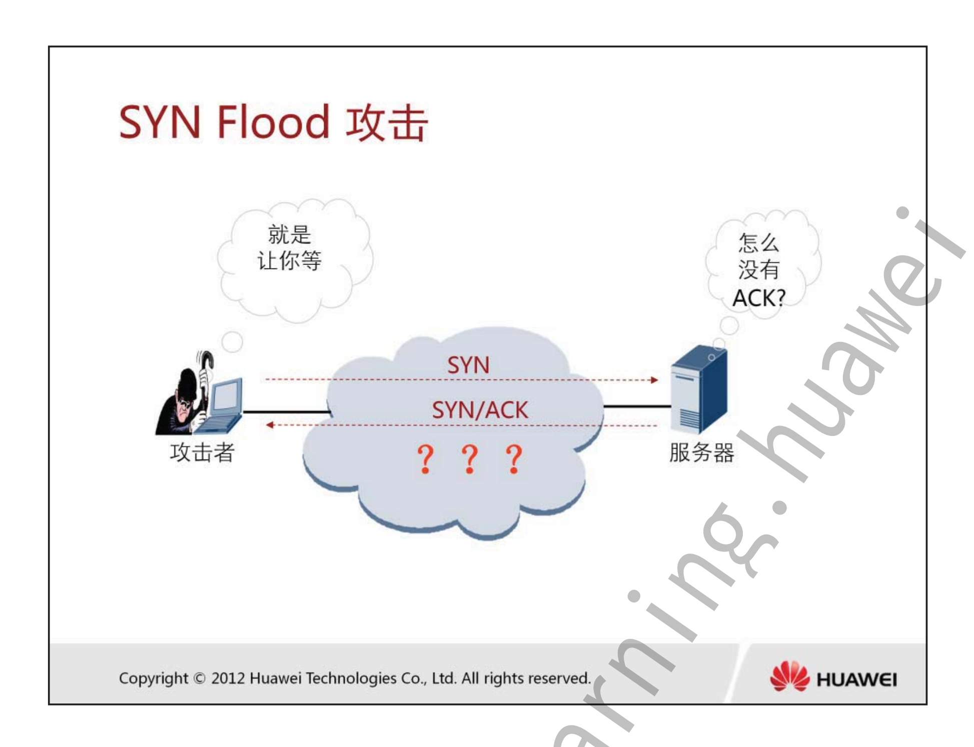
检测方法：

WinNuke攻击1检测数据包目标端口是否为139、138、137等，并判断URG位是否为“1”且URG指针不为空。

WinNuke攻击2 检测进入的IGMP报文是否为分片报文，如是分片报文，则直接丢弃。

配置：

- firewall defend winnuke enable



攻击介绍：

- 由于资源的限制，TCP/IP栈的实现只能允许有限个TCP连接。而SYN Flood攻击正是利用这一点，它伪造一个SYN报文，其源地址是伪造的、或者一个不存在的地址，向服务器发起连接，服务器在收到报文后用SYN-ACK应答，而此应答发出去后，不会收到ACK报文，造成一个半连接。如果攻击者发送大量这样的报文，会在被攻击主机上出现大量的半连接，耗尽其资源，使正常的用户无法访问。直到半连接超时。在一些创建连接不受限制的环境里，SYN Flood具有类似的影响，它会消耗掉系统的内存等资源。

处理方法：

- 对TCP连接数进行统计，可以进行基于IP地址或区域的统计，当TCP的每秒新建连接速率超过指定的阈值或超过规定的总数目时，认为存在SYN Flood攻击，可以采用TCP代理技术或者采用新建连接的限制或总连接数的限制。

备注：对于这种需要进行统计分析的攻击防范，需要开启防火墙的统计功能。

SYN Flood 攻击（续）

配置

- firewall defend syn-flood interface { *interface-type interface-number* | all } [alert-rate *alert-rate-number1*] [max-rate *max-rate-number1*] [tcp-proxy { auto | off | on }]
- firewall defend syn-flood zone [vpn-instance *vpn-instance-name*] *zone-name* [alert-rate *alert-rate-number2*] [max-rate *max-rate-number2*] [tcp-proxy { auto | on | off }]
- firewall defend syn-flood enable



interface-type interface-number ： 接口类型和接口编号。

all： 所有的接口。表示开启基于所有接口的SYN Flood攻击防范功能
。

alert-rate-number1： 告警速率。 整数形式，取值范围为1～1000000，单位为pps。缺省值为16000pps。

max-rate-number1： 最大速率。 整数形式，取值范围为1～1000000，单位为pps。缺省值为500000pps，最大速率的值必须大于告警速率。

tcp-proxy： TCP代理功能。 缺省情况下，TCP代理为自动开启状态。当tcp-proxy的参数设置为auto时，TCP代理的启动状态为自动开启，即连接数超过告警速率时开启；当参数设置为on时，所有连接均使用代理连接；当参数设置为off时，不开启TCP代理功能。

alert-rate-number2： 告警速率。 整数形式，取值范围为1～65535，单位为cps。缺省值为1000cps。

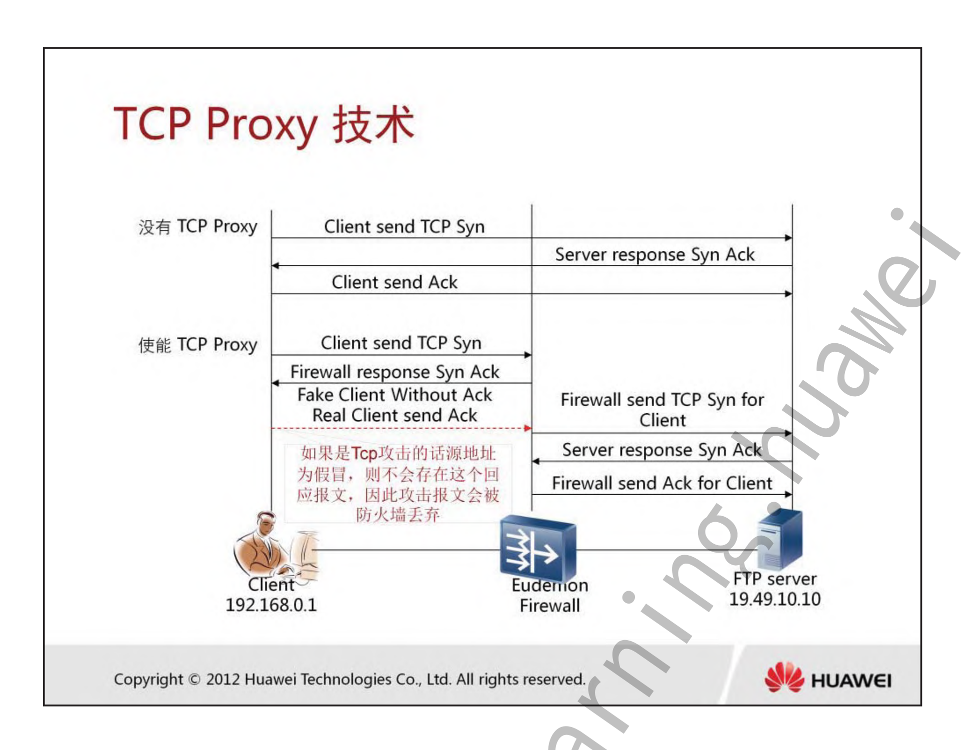
vpn-instance-name： VPN实例的名称。 必须为已经创建的VPN实例名称。

max-rate-number2： 最大速率。 整数形式，取值范围为1～65535，单位为cps。缺省值为50000cps。最大速率的值必须大于告警速率。

zone-name： 安全区域的名称。 必须为已经存在安全区域名称。

说明:

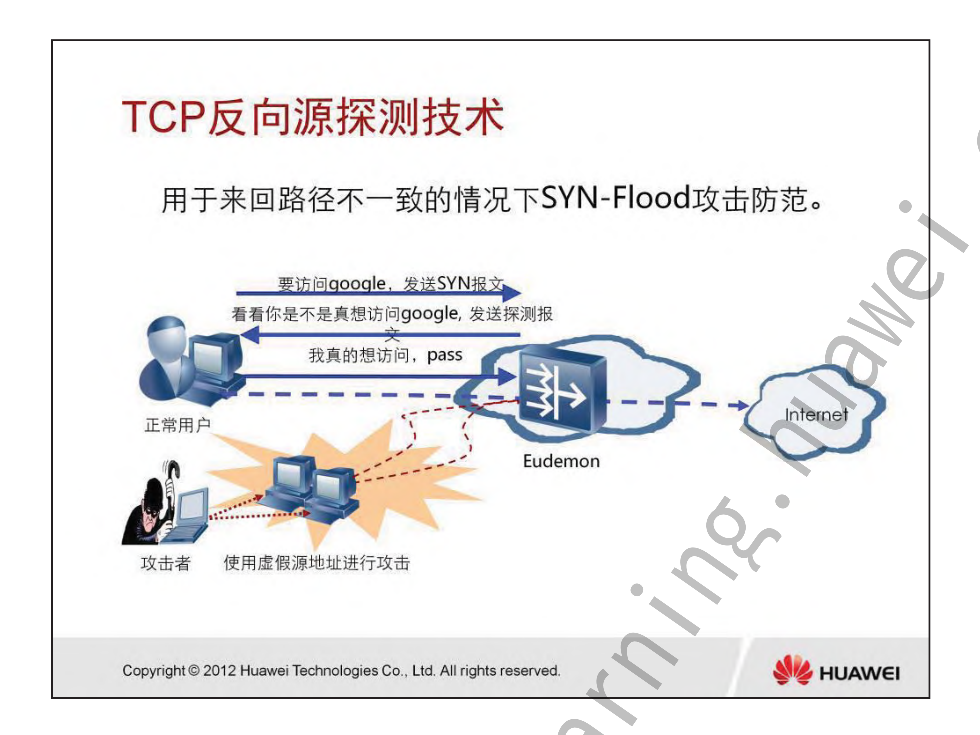
当设备开启SYN Flood攻击防范TCP代理功能，而网络中设备接口存在较小MTU时，请设置TCP-MSS值；如果MSS设置过大，TCP连接上传输的报文较大而又不允许分片时，报文可能会被网络中的设备丢弃，部分tcp业务将受到影响



TCP代理：

- 防止SYN Flood攻击的一个有效的办法就是采用防火墙的TCP代理功能。 我们把连接发起端称为客户，对端称为服务器，它们通过防火墙的中继进行通信。客户发起连接，防火墙并不把SYN 包传递给服务器，而是自己伪装成服务器返回应答；客户确认后再以当初客户发起连接时的信息向服务器发起连接。当客户和服务器之间传输的数据通过防火墙时，防火墙只需对它们的序号进行调整就可以了。
- 通过TCP代理功能，防火墙就能拦截所有到达的连接请求，并代表服务器建立与客户机的连接，代表客户机建立与服务器的连接。如果两个连接都成功地建立，防火墙就会将两个连接进行中继。这样防火墙就能很好的保护服务器不受SYN-Flood的攻击，同时防火墙有更严格的超时限制，以防止其自身的资源被 SYN 攻击耗尽。
- 同时也可以采用针对半开连接的数目和速率等来监控syn flood攻击，检测特定目的地址SYN报文的接收速率和TCP半开连接数。当特定目的地址SYN报文的连接速率或TCP半连接数超过设定的阈值时，通知系统目的主机受到SYN Flood攻击，并根据攻击防范配置决定是否启动TCP代理功能；

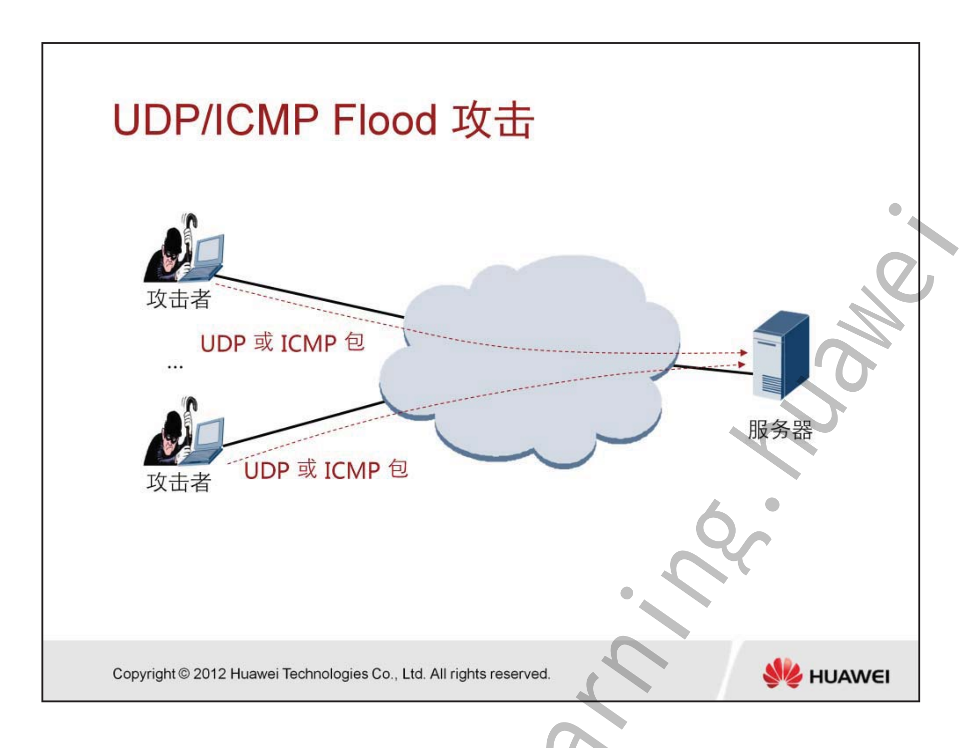
- 当SYN报文连接速率和TCP半连接数均低于阈值的80%时，通知系统SYN Flood攻击结束，并根据攻击防范配置决定是否关闭代理功能。当攻击发生时，对攻击记录日志。



第一步： 通过响应报文的校验确认源是否合法， 可以防止非法虚假源攻击；

第二步： 源若合法，通过TTL跳数确认是否是假冒其他合法地址发起的攻击。

TCP反向源探测会在client端发送SYN报文后，由防火墙首先回应一个ack_sequence序号错误的SYN-ACK报文，此报文的ack_sequence，为HASH值，其中包含此报文的TTL，与SYN报文的sequence+1不一致，如果client端真实存在收到此SYN-ACK就会发送RST报文，防火墙收到此RST报文，判断sequence与上述syn-ack的ack_sequence是否一致，如果一致，从此序号恢复TTL，与报文的TTL比较，如果相等或者相差很小范围，则认为SYN及RST报文都是从同一地址发出，此源IP加为认证通过状态。后续再有SYN报文到达，命中源IP监控表，发现认证通过，则转发此SYN报文。



攻击介绍：

- 短时间内向特定目标发送大量的UDP/ICMP报文，致使目标系统负担过重而不能处理合法的连接。一般这种攻击以分布式居多，流量大。

处理方法：

- 检测通向特定目的地址的UDP报文的速率，当速度超过设定的阈值上限时，设定攻击标志并做Car处理，对攻击记录日志。当速率低于设定的阈值下限，取消攻击标志，允许所有报文通向特定目的地址。

使用限制：

- 无使用限制，注意配置正确的目的保护地址。ICMP/UDP是无连接的协议，因此不能提供类似SYN FLOOD代理方式的防御方法。

UDP/ICMP Flood攻击（续）

配置

- firewall defend udp-flood interface { *interface-type interface-number* | all } [max-rate *max-rate-number1*]
- firewall defend udp-flood zone [vpn-instance *vpn-instance-name*] *zone-name* [**alert-rate** *alert-rate-number*] [max-rate *max-rate-number2*]
- **firewall defend icmp-flood interface** { *interface-type interface-number* | **all** } [**max-rate** *max-rate-number1*]
- **firewall defend icmp-flood zone** [**vpn-instance** *vpn-instance-name*] *zone-name* [**max-rate** *max-rate-number2*]
- firewall defend udp/icmp-flood enable



interface-type interface-number :接口类型和接口编号。

all :所有的接口。

max-rate-number1 :最大速率。 整数形式，取值范围为1～500000，单位为pps。缺省为500000pps。

vpn-instance-name :VPN实例的名称。 必须为已经创建的VPN实例名称。

max-rate-number2 :最大速率。 整数形式，取值范围为1～65535，单位为cps。缺省为1000cps。

zone-name 安全区域的名称。 必须为已经存在的安全区域名称。

说明：

配置的UDP/ICMP-flood攻击防范参数在开启攻击防范功能之后才生效。

其它Flood攻击手段

- DNS Flood
- Get Flood
- Tcp-illeage-session



DNS flood 用来对DNS服务器发起flood请求

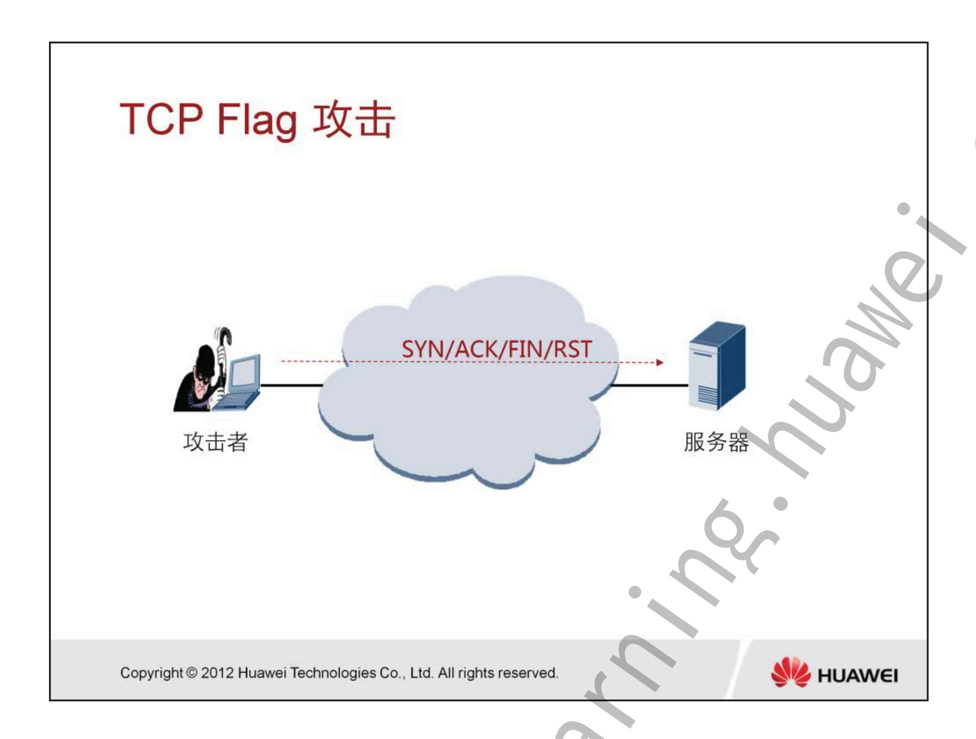
Get flood 用来对WEB 服务器发起Get 操作请求

TCP-illeage-session用来对服务器发起TCP空连接请求，耗尽服务器的连接资源。



目 录





攻击介绍：

TCP报文包含6个标志位：URG、ACK、PSH、RST、SYN、FIN，不同的系统对这些标志位组合的应答是不同的：

- 6个标志全部为1，也就是圣诞树攻击；
- 6个标志全部为0，如果端口是关闭的，会使接收方应答一个RST | ACK消息。而对于一个开放端口，Linux和UNIX机器不会应答，而Windows机器将回答RST | ACK消息。这可用于操作系统探测。
- 不管端口是打开还是关闭，ACK与除RST外的其它任何一个状态位组合在一起，都会引起一个还没有发送请求的接收方的一个RST应答。这可用于探测主机的存在。
- 不管端口是打开还是关闭，SYN | FIN | URG 会让接收方发送一个 RST | ACK 应答，这可用于探测主机的存在。
- 如果端口是关闭的，SYN、SYN | FIN、SYN | PUSH、SYN | FIN | PUSH、SYN|URG、SYN | URG | PUSH、SYN | FIN | URG | PUSH 会使接收方应答一个RST | ACK消息；如果端口是打开的，会使接收方应答一个 SYN | ACK消息，这可用于主机探测和端口探测。

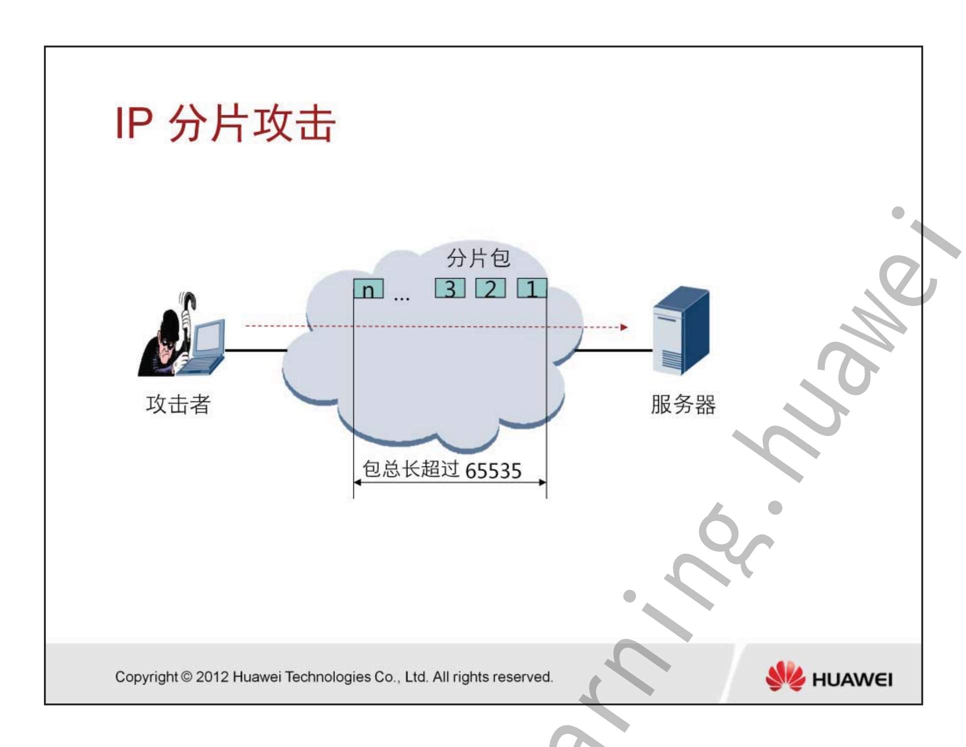
- 如果端口是关闭的，FIN 、URG、PUSH、URG|FIN 、URG|PUSH 、FIN|PUSH 、URG|FIN|PUSH 会使接收方应答一个RST | ACK消息。而对于一个开放端口，Linux和UNIX机器不会应答，而Windows机器将回答RST | ACK消息。这可用于操作系统探测。

处理方法：

- 检查TCP报文的各个标志位，若出现
- 6个标志位全为1；
- 6个标志位全为0；
- SYN和FIN位同时为1；
- 直接丢弃满足以上任一条件的报文，并记录日志。

配置：

- firewall defend tcp-flag enable



攻击介绍：

IP报文中有几个字段与分片有关：DF位、MF位，Fragment Offset 、Length 。

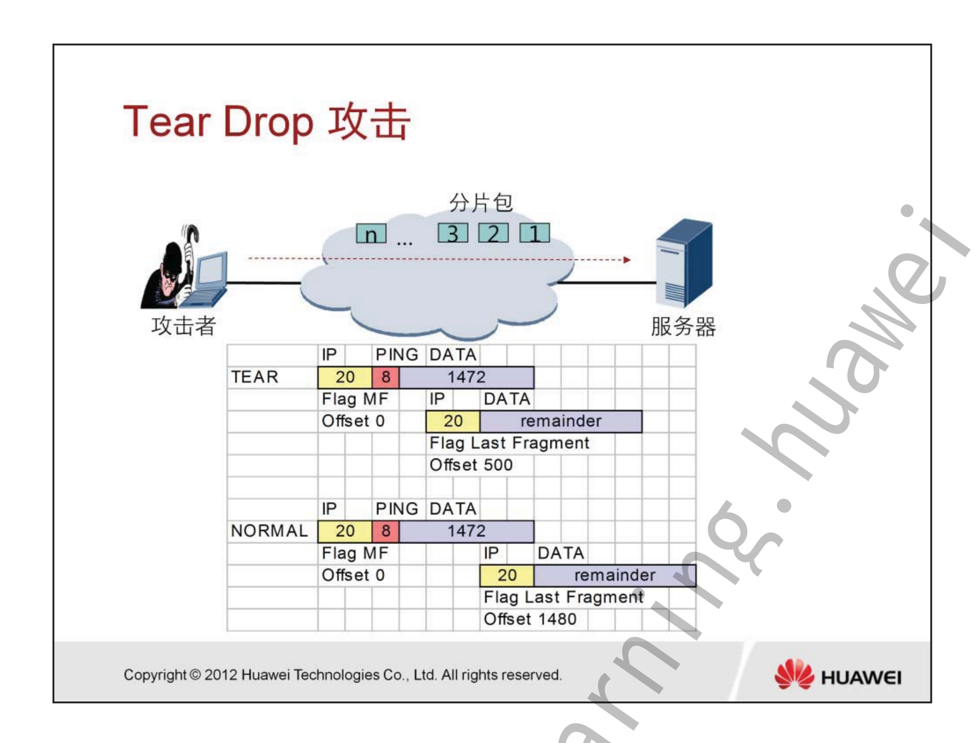
- 如果上述字段的值出现矛盾，而设备处理不当，会对设备造成一定的影响，甚至瘫痪。
- 矛盾的情况有：
- DF位被置位，而MF位同时被置位或Fragment Offset不为0；
- DF位为0，而Fragment Offset + Length > 65535。

处理方法：

- 若分片报文的目的地址为本防火墙，则直接丢弃；
- 检查IP报文中与分片有关的字段（DF位、MF位、片偏置量、总长度）是以下问题：
  - DF位为1，且MF位也为1。
  - DF位为1，且Offset > 0。
  - DF位为0，且分片 Offset + Length > 65535。

配置：

- firewall defend ip-fragment enable



攻击介绍：

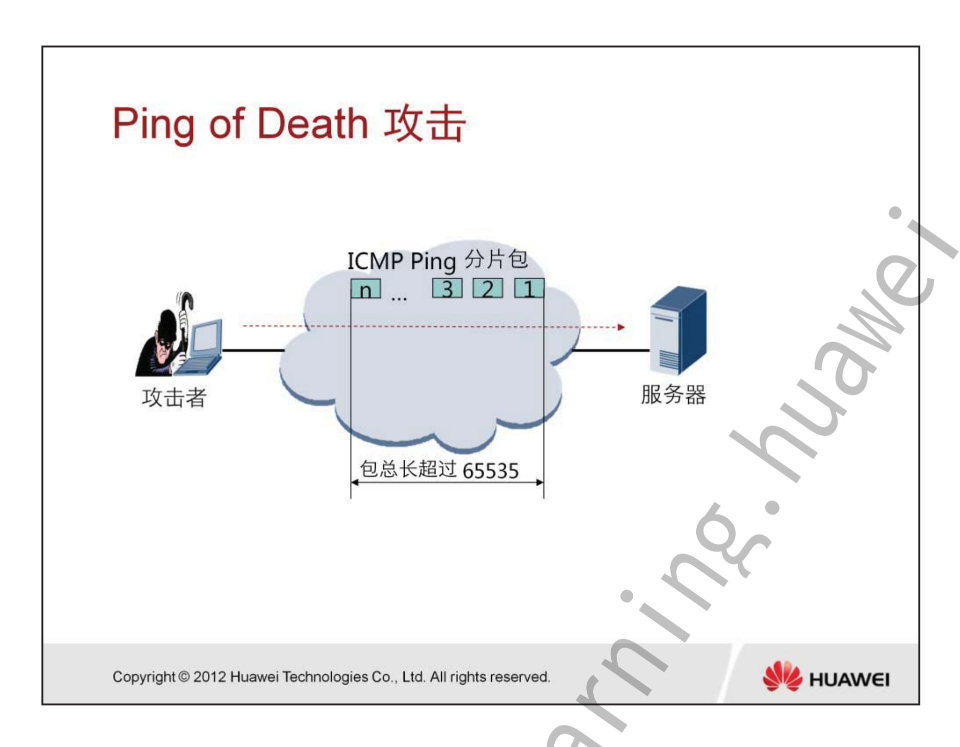
- Teardrop攻击是指攻击者截取IP数据包后，把偏移字段设置成不正确的值，接收端在收到这些分拆的数据包后，就不能按数据包中的偏移字段值正确组合出被拆分的数据包，这样，接收端会不停的尝试，以至操作系统因资源耗尽而崩溃。

处理方法：

- 缓存分片信息，每一个源地址、目的地址、分片ID相同的为一组，最大支持缓存10000组分片信息。在分片缓存的组数达到最大时，如果后续分片报文要求建立新组，则直接丢弃。

配置：

- firewall defend teardrop enable



攻击介绍：

- IP报文的长度字段为16位，这表明一个IP报文的最大长度为65535。对于ICMP ECHO Request报文，如果数据长度大于65508，就会使ICMP数据 + IP头长度(20) + ICMP头长度（8）> 65535。对于有些路由器或系统，在接收到一个这样的报文后，由于处理不当，会造成系统崩溃、死机或重启。 所谓Ping of Death，就是利用一些尺寸超大的ICMP报文对系统进行的一种攻击。

处理方法：

- 检测ICMP回送请求报文的长度是否超过65535字节，若超过，则丢弃报文，并记录日志。

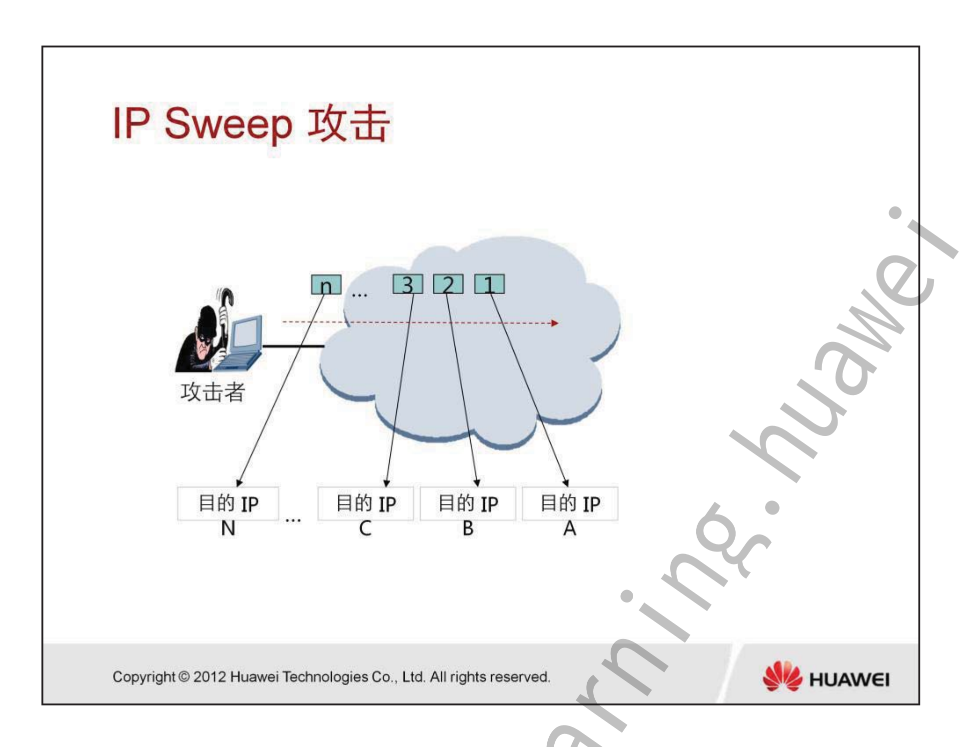
配置：

- firewall defend ping-of-death enable



目 录





攻击介绍：

- 运用ping这样的程序探测目标地址，对此作出响应的表示其存在，用来确定哪些目标系统确实存活着并且连接在目标网络上。也有可能使用TCP/UDP报文对某些地址发起连接（如TCP ping），判断是否有应答报文。

处理方法：

- 检测进入防火墙的ICMP、TCP和UDP报文，由该报文的源IP地址获取统计表项的索引，如目的IP地址与前一报文的IP地址不同，则将表项中的总报文个数增1。如果在一定时间内报文的个数达到设置的阈值，直接丢弃报文，记录日志，并根据配置决定是否将源IP地址自动加入黑名单。被加入黑名单的IP，其报文将不能通过防火墙。当然，如果某些IP被发现有恶意攻击或占用过量带宽，也可以手工把该IP加入黑名单。

使用限制：

- 扫描类攻击的源地址是真实的，因此可以采用直接加入黑名单的方法进行防御。
- 扫描类攻击的扫描速度决定了攻击防范的有效性。
- 蠕虫病毒爆发的时候，一般就是地址扫描攻击。

IP Sweep 攻击（续）

配置:

- **firewall defend ip-sweep** { **max-rate** *rate-number* | **blacklist-timeout** *interval* }
- **Firewall blacklist enable**



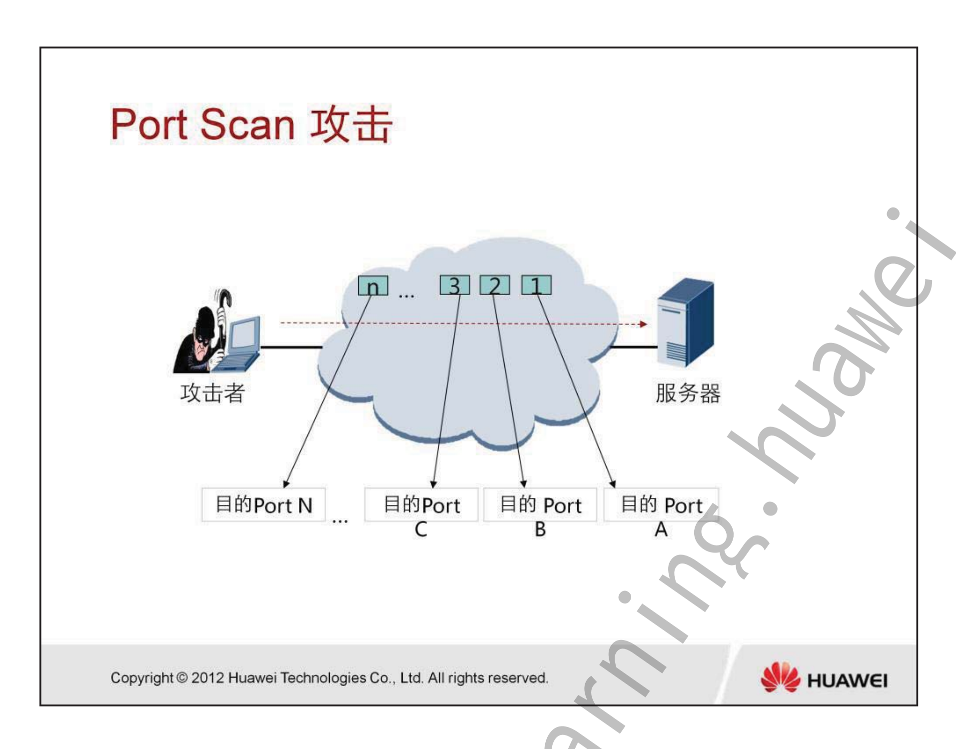
max-rate rate-number：设定从同一源地址向外发送报文的目的地址变化速率的阈值。

rate-number 取值范围是1 ～ 10,000，单位是次/秒。默认值为4000次/秒。

blacklist-timeout interval：将攻击源IP 加入黑名单并设定其在黑名单内的保持时间，

interval 取值范围是1 ～ 1000，单位分钟，默认值为20。

注意：如果要启用黑名单隔离功能，需要先启动黑名单。



攻击介绍：

- Port Scan攻击通常使用一些软件，向大范围的主机的一系列TCP/UDP端口发起连接，根据应答报文判断主机是否使用这些端口提供服务。

处理方法：

- 检测进入防火墙的TCP报文或UDP报文，由该报文的源IP地址获取统计表项的索引，如目的端口与前一报文不同，将表项中的报文个数增1。如果报文的个数超过设置的阈值，直接丢弃报文，记录日志，并根据配置决定是否将源IP地址加入黑名单。

使用限制：

- 扫描类攻击的源地址是真实的，因此可以采用直接加入黑名单的方法进行防御。
- 扫描类攻击的扫描速度决定了攻击防范的有效性。

Port Scan 攻击（续）

配置

- **firewall defend port-scan** [**max-rate** *rate-number*] [**blacklist-timeout** *interval*]
- **Firewall blacklist enable**



max-rate rate-number：设定从同一源地址向外发送报文的目的端口变化速率的阈值。

rate-number 取值范围是1 ～ 10,000，单位是次/秒。默认值为4000次/秒。

blacklist-timeout interval：将攻击源IP 加入黑名单并设定其在黑名单内的保持时间。

interval 取值范围是1 ～ 1000，单位是分钟。默认值为20 分钟。

注意：如果要启用黑名单隔离功能，需要先启动黑名单。

防火墙防范的其他报文

ICMP Redirect

ICMP Unreachable

Large ICMP

Route Record

Tracert



ICMP重定向报文

攻击介绍：

网络设备向同一个子网的主机发送ICMP重定向报文，请求主机改变路由。一般情况下，设备仅向主机而不向其它设备发送ICMP重定向报文。但一些恶意的攻击可能跨越网段向另外一个网络的主机发送虚假的重定向报文，以期改变主机的路由表，干扰主机正常的IP报文转发。

处理方法：

根据本控制功能的使能状态对ICMP重定向报文（类型为5）进行转发或丢弃。在禁止转发时，如发现此类报文到达，则记录日志。除记录日志模块规定的内容外，还记录重定向到何处。

ICMP不可达报文

攻击介绍：

不同的系统对ICMP不可达报文（类型为3）的处理不同，有的系统在收到网络（代码为0）或主机（代码为1）不可达的ICMP报文后，对于后续发往此目的地的报文直接认为不可达，好像切断了目的地与主机的连接，造成攻击。

处理方法：

根据本控制功能的使能状态对类型号为3的ICMP不可达报文进行转发或丢弃。在禁止转发时，如发现此类报文到达，则记录日志。

超大的ICMP报文

攻击介绍：

一般来说，网络中传输的ICMP报文都不大，而且对于不同的系统，能够发送和接收的最大ICMP报文的大小也是不一样的，因此出现超大的ICMP报文，应为异常现象。

处理方法：

检测ICMP报文长度是否超过设定的最大值，如果超过，则直接丢弃，并记录日志。

IP路由记录选项

攻击介绍：

同IP源站选路功能类似，在IP 路由技术中，还提供了路由记录选项。它的含义记录IP报文从源到目的过程中所经过的路径，也就是一个处理过此报文的路由器的列表。IP路由记录选项通常用于网络路径的故障诊断，但也会被恶意攻击者利用，刺探网络结构。

处理方法：

检测进入路由器的报文是否设置IP路由记录选项，如是，则丢弃报文，并记录日志。

Tracert报文

攻击介绍：

Tracert是利用TTL为0时返回的ICMP超时报文，和达到目的地时返回的ICMP端口不可达报文来发现报文到达目的地所经过的路径，它可以窥探网络的结构。

处理方法：

检查ICMP报文是否为超时（类型为11）或为目的端口不可达报文（类型号为3，代码号为3），如果是，则根据本控制功能的使能状态选择对报文进行转发或丢弃。在禁止转发时，如发现此类报文到达，则记录日志。



问 题

如何手工绑定PC的网关MAC地址，防止ARP欺骗劫持流量?



答案:

- 如果是在windows的PC主机上，通过命令： arp -s X.X.X.X XX-XX-XX-XX-XX-XX，静态绑定网关的IP地址和MAC地址，但须注意，PC一旦重启的话，刚绑定配置会丢失，需要重新绑定。

谢谢
www.huawei.com


USG防火墙
双机热备业务特性与配置
www.huawei.com
Copyright © 2012 Huawei Technologies Co., Ltd. All rights reserved.
HUAWEI



前 言

在当前的组网应用中，用户对网络可靠性的要求越来越高，特别是在一些重要的业务入口或接入点上需要保证网络不间断运行。对于这些重要的业务点如何保证网络的不间断传输，成为必须解决的一个问题。

本胶片主要介绍防火墙的双机热备份技术原理和具体配置，以及在USG防火墙上实施双机热备份技术所使用的三种协议：VRRP、VGMP和HRP。





培训目标

学完本课程后，您应该能：

- 掌握双机热备份技术原理
- 掌握VRRP，VGMP和HRP之间的关系
- 掌握典型双机组网的配置

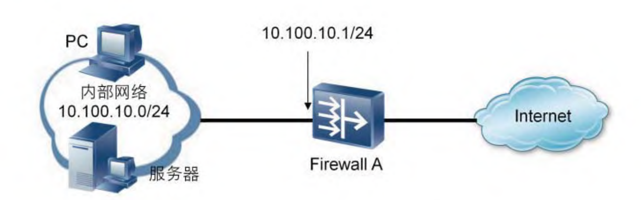




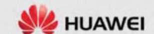
目 录



双机热备份技术产生的原因

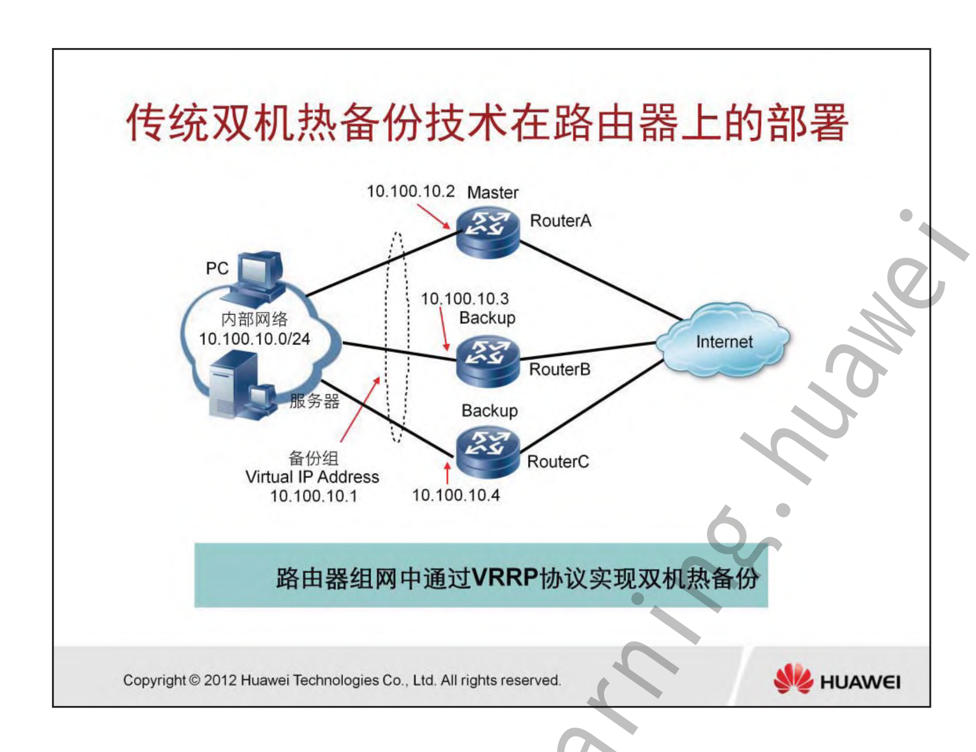


传统的组网方式如图所示，内部用户和外部用户的交互报文全部通过Firewall A。如果Firewall A出现故障，内部网络中所有以Firewall A作为默认网关的主机与外部网络之间的通讯将中断，通讯可靠性无法保证。



双机热备份技术的出现改变了可靠性难以保证的尴尬状态，通过在网络出口位置部署两台或多台网关设备，保证了内部网络与外部网络之间的通讯畅通。

USG防火墙作为安全设备，一般会部署在需要保护的网络和不受保护的网络之间，即位于业务接口点上。在这种业务点上，如果仅仅使用一台USG防火墙设备，无论其可靠性多高，系统都可能会承受因为单点故障而导致网络中断的风险。为了防止一台设备出现意外故障而导致网络业务中断，可以采用两台防火墙形成双机备份。

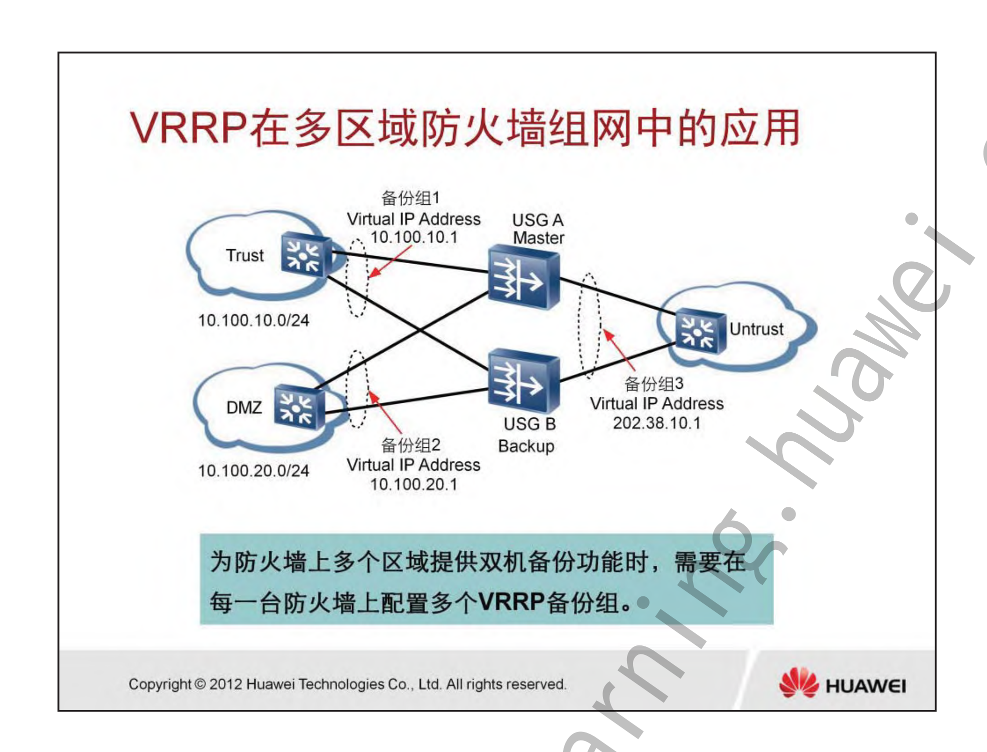


为了避免路由器传统组网所引起的单点故障的发生，通常情况可以采用多条链路的保护机制，依靠动态路由协议进行链路切换。但这种由路由协议来进行切换保护的方式存在一定的局限性，当不能使用动态路由协议时，仍然会导致链路中断的问题，因此推出了另一种保护机制VRRP（虚拟路由冗余协议）。采用VRRP的链路保护机制比依赖动态路由协议的广播报文来进行链路切换的时间更短，同时弥补了不能使用动态路由情况下的链路保护。

VRRP（Virtual Router Redundancy Protocol）是一种基本的容错协议，它将同一个广播域的一组路由器组织成一个虚拟路由器，称之为一个备份组。其中仅有一台设备处于活动状态（称为主设备（Master）），其余设备都处于备份状态（称为备份设备（Backup））。每个备份组（即虚拟路由器）拥有一个虚拟IP地址。只有处于活动状态的路由器能转发以虚拟IP地址作为下一跳的报文。

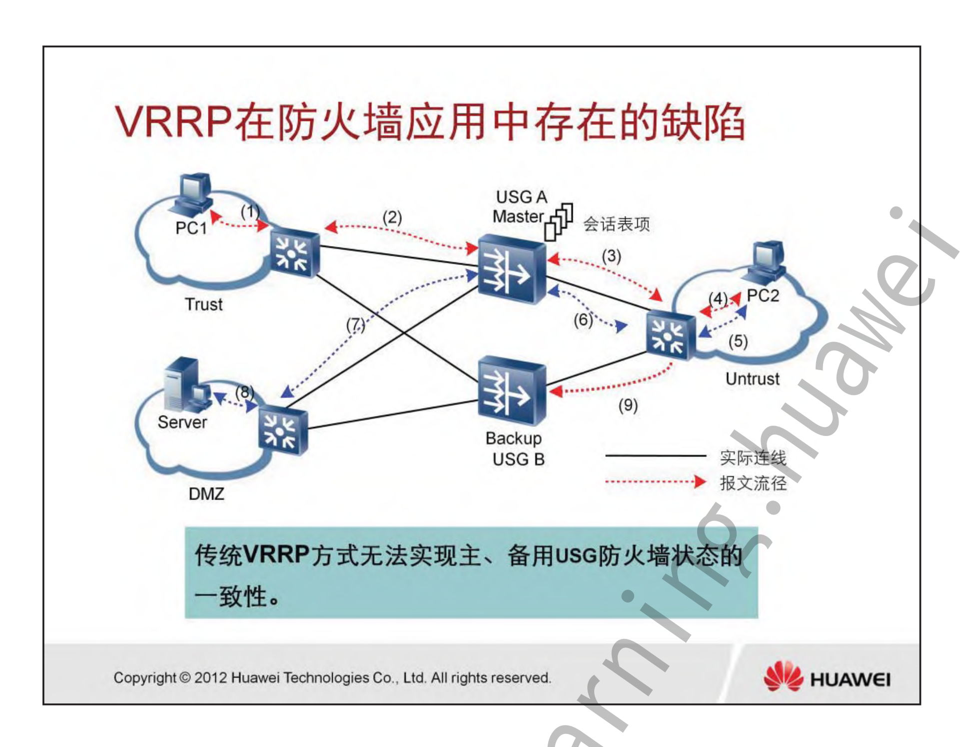
内部网络中的所有主机仅仅知道该虚拟IP地址，而并不知道具体的主用（Master）或备用（Backup）设备的IP地址，因此各主机都将缺省路由配置为去往该虚拟IP地址。于是，内部网络中的各主机就通过该备份组与外部网络进行通信。

Master路由器VRRP协议模块监视通信接口状态，并通过组播方式向Backup路由器发送通告报文。如果Master路由器接口故障或链路出现问题，则会导致无法正常发送VRRP通告报文。当Backup路由器在指定时间内收不到VRRP通告报文时，VRRP协议就把Backup设备的VRRP状态变成主用，并且将转发以虚拟路由器IP地址作为下一跳的报文，从而将相关通信切换到该设备（原先的备份设备）上。因此，采用VRRP技术实现了内部网络中的主机不间断地与外部网络进行通信，可靠性得到保证。



由于USG防火墙是状态防火墙，它要求报文的来回路径通过同一台防火墙。为了满足这个限制条件，就要求在同一台防火墙上的所有VRRP备份组状态保持一致，即要保证在某一台防火墙上所有VRRP备份组都是主状态，这样所有报文都将从此防火墙上通过，而另外一台防火墙则充当备份设备。

当防火墙上多个区域需要提供双机备份功能时，需要在一台防火墙上配置多个VRRP备份组。



如图所示，假设USG A和USG B的VRRP状态一致，即USG A的所有接口均为主用状态，USG B的所有接口均为备用状态。

- 此时，Trust区域的PC1访问Untrust区域的PC2，报文的转发路线为(1)-(2)-(3)-(4)。USG A转发访问报文时，动态生成会话表项。当PC2的返回报文经过(4)-(3)到达USG A时，由于能够命中会话表项，才能再经过(2)-(1)到达PC1，顺利返回。同理，PC2和DMZ区域的Server也能互访。
- 假设USG A和USG B的VRRP状态不一致，例如，当USG B与Trust区域相连的接口为备用状态，但与Untrust区域的接口为主用状态，则PC1的报文通过USG A设备到达PC2后，在USG A上动态生成会话表项。PC2的返回报文通过路线(4)-(9)返回。此时由于USG B上没有相应数据流的会话表项，在没有其他报文过滤规则允许通过的情况下，USG B将丢弃该报文，导致会话中断。

防火墙备份技术产生的原因：

- 报文的转发机制不同：
- 对于路由器来说，业务中的每个数据包都会逐包转发，即每个报文都会查路由表，当匹配上后才进行转发。

- 当链路切换后，后续报文不会受到影响，继续进行转发。而由于USG是状态防火墙，只会对业务中首包进行检查，如果首包允许通过会建立一条五元组的会话连接，只有命中该会话表项的后续报文（包括返回报文）才能够通过USG防火墙，如果链路切换后，后续报文找不到正确的表项，会导致业务中断。其实路由器在配置NAT后也会存在同样的问题，因为在进行NAT后会形成一个NAT转换后的表项。
- VRRP在防火墙应用的缺陷:
    - USG连接多个安全区域，和每个安全区域相关的接口均形成一个备份组，按照传统的VRRP机制，VRRP均为相对独立，且单独工作的。由此，无法保证同一防火墙上各接口的VRRP状态都为主用或都为备用，即传统VRRP方式将无法实现USG VRRP状态的一致性。即使VRRP状态一致，如果发生状态切换，主用防火墙上生成的会话表不会备份到备用防火墙上，同样会导致业务中断。
    - 所谓双机热备其实是双机状态备份，当两台防火墙，在确定主从防火墙后，由主防火墙进行业务的转发，而从防火墙处于监控状态，同时主防火墙会定时向从防火墙发送状态信息和需要备份的信息，当主防火墙出现故障后，从防火墙会及时接替主防火墙上的业务运行。



目 录



防火墙双机热备份技术分析

防火墙双击热备份技术的特征

- 控制主、备用防火墙的切换
- 状态信息的备份

USG防火墙的双机热备份技术依靠三种协议实现：

- VRRP（虚拟路由冗余协议）
- VGMP（VRRP组管理协议）
- HRP（华为冗余协议）



防火墙双机热备份技术需要完成两项工作：

- 控制主、备用防火墙的切换
- 状态信息的备份

以上两项工作需要由VGMP和HRP配合完成，而基本的热备份功能则由VRRP实现。

VGMP：VRRP Group Management Protocol

HRP： Huawei Redundancy Protocol

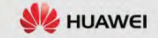
防火墙主备状态切换的实现

VGMP（VRRP Group Management Protocol) 提出VRRP管理组的概念，将同一台防火墙上的多个VRRP备份组都加入到一个VRRP管理组，由管理组统一管理所有VRRP备份组。通过统一控制各VRRP备份组状态的切换，来保证管理组内的所有VRRP备份组状态都是一致的。

VGMP的作用：防火墙主备状态控制切换

VRRP管理组的功能：

- 状态一致性管理（管理组内VRRP备份组同步状态切换）
- 抢占管理（屏蔽VRRP备份组抢占）
- 通道管理（trans-only）



状态一致性管理

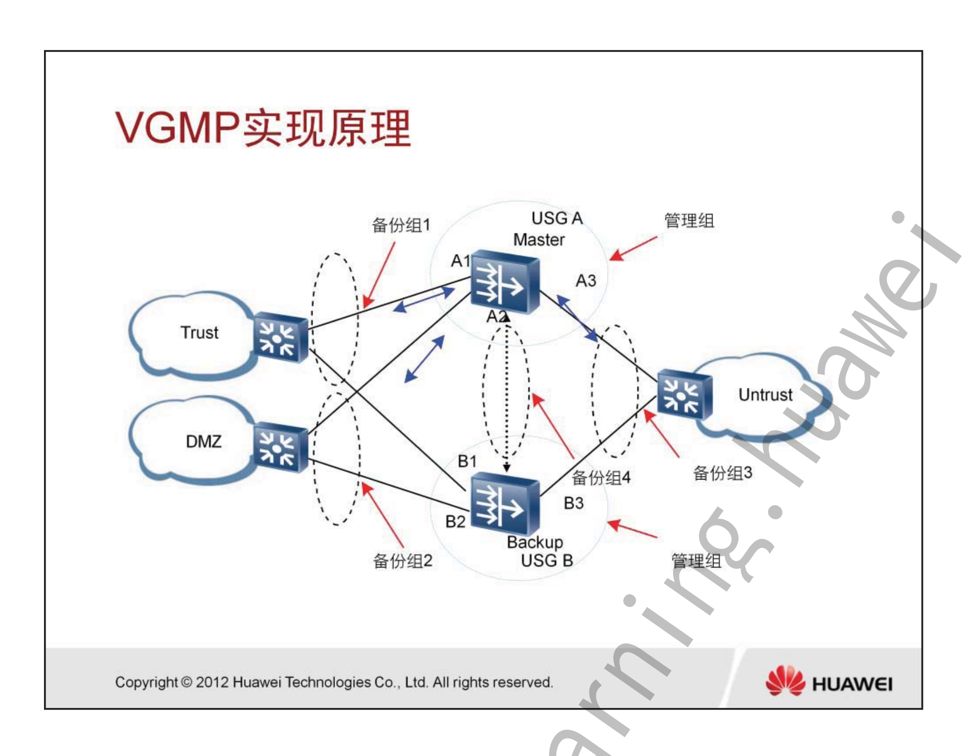
- 各备份组的主/备状态变化都需要通知其所属的VGMP管理组，由VGMP管理组决定是否允许VRRP备份组进行主/备状态切换。如果需要切换，则VGMP管理组控制所有的VRRP备份组统一切换。VRRP备份组加入到管理组后，状态不能自行单独切换。

抢占管理

- VRRP备份组本身具有抢占功能。即当原来出现故障的主设备故障恢复时，其优先级也会恢复，此时可以重新将自己的状态抢占为主。
- VGMP管理组的抢占功能和VRRP备份组类似，当管理组中出现故障的备份组故障恢复时，管理组的优先级也将恢复。此时VGMP可以决定是否需要重新抢占称为主设备。
- 当VRRP备份组加入到VGMP管理组后，备份组上原来的抢占功能将失效，抢占行为发生与否必须由VGMP管理组统一决定。

通道管理

- 通道是双机热备的防火墙之间用来传输VGMP、HRP报文的一对VRRP备份组。所谓通道管理，就是为了确定双机热备的两台防火墙之间有哪些通道是可用的，VGMP、HRP模块将自动选用可用的传输通道来发送VGMP、HRP报文。
- 通道管理的实现依赖于主备防火墙的VGMP之间定时发送的Hello报文，其中携带了双方的各个备份组成员的状态信息，由此可以判断通道的可用性。

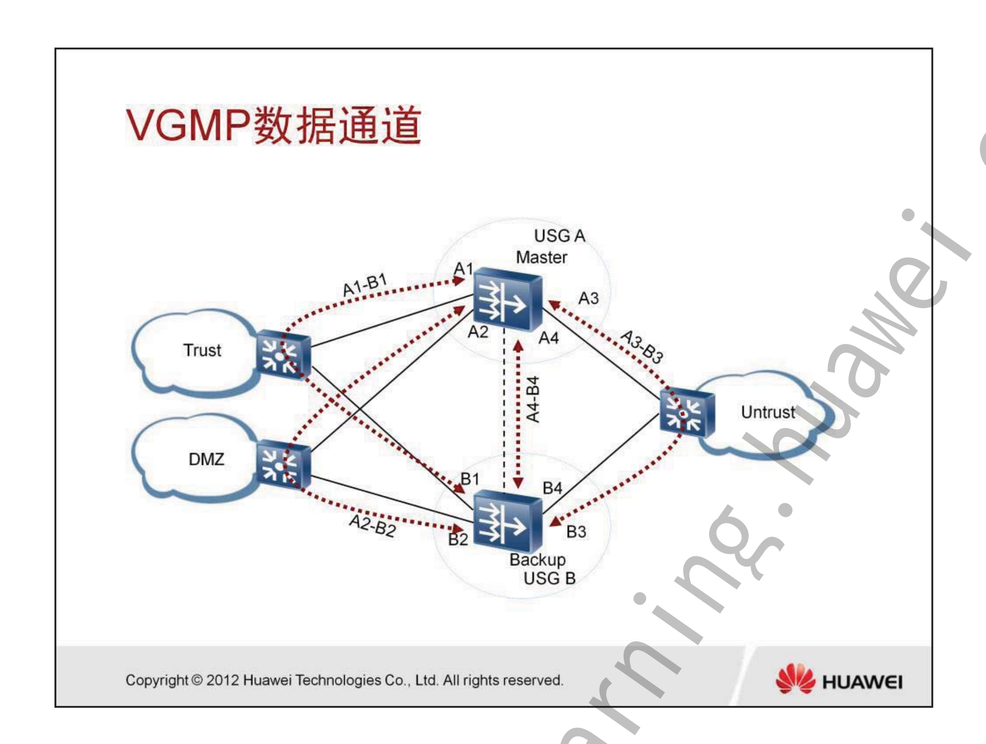


VGMP管理组状态(Master/Slave)

- 当防火墙上的VGMP管理组为Master状态时，它保证组内所有VRRP备份组的状态统一为Master状态，这样所有报文都将从该防火墙上通过，该防火墙成为主用防火墙。此时另外一台防火墙上对应的VGMP管理组为备状态，该防火墙成为备用防火墙。
- 在配置VGMP时，也需要指定VGMP管理组的优先级，两台防火墙根据优先级决定谁将成为主防火墙。VGMP管理组的优先级会根据组内的VRRP备份组成员的状态动态调整，以此完成两台防火墙的主备倒换。

VGMP管理组主、备状态的通告报文（Hello）

- 与VRRP类似，状态为Master的VGMP也会定期向对端发送Hello报文，通知Slave端本身的运行状态（包括优先级、VRRP成员状态等）。与VRRP不同的是，Slave端收到Hello报文后，会回应一个ACK消息，该消息中也会携带本身的优先级、VRRP成员状态等。两台防火墙通过Hello报文交互各自的状态信息。
- VGMP Hello报文发送周期缺省为1000ms。当Slave端三个Hello报文周期没有收到对端发送的Hello报文时，会认为对端出现故障，从而将自己切换到Master状态。



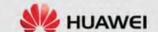
两台防火墙的VGMP管理组之间通信的VGMP报文、数据备份的HRP报文，都是通过VGMP管理组中的数据通道进行传输的。数据通道是指两端防火墙的VGMP管理组中相对应的一对VRRP备份组成员。如下图中共有四个数据通道。

两台防火墙之间的直连的链路（如图中的A4-B4），一般也称为HRP心跳线。使用专用的链路来承载VGMP和HRP报文可以提高双机热备的可靠性。

防火墙状态信息的备份

VGMP可以保证报文来回路径通过同一台防火墙。当主防火墙出现故障时，所有流量都将切换到备防火墙。但USG防火墙是状态防火墙，如果备防火墙上没有原来主防火墙上的连接状态数据，则切换到备防火墙的很多流量将无法通过，造成现有的连接中断，此时用户必须重新发起连接。

为了实现主用设备出现故障时能由备用设备平滑地接替工作，需要在主、备用设备之间备份关键配置命令和会话表状态等关键信息。



HRP的出现，解决了主备用防火墙状态信息备份的难题。

HRP

HRP（Huawei Redundancy Protocol ）华为冗余协议

华为公司冗余协议HRP（Huawei Redundancy Protocol）是承载在VGMP报文上进行传输的。HRP用于在主用设备和备用设备之间备份关键配置命令和会话表状态等关键信息。



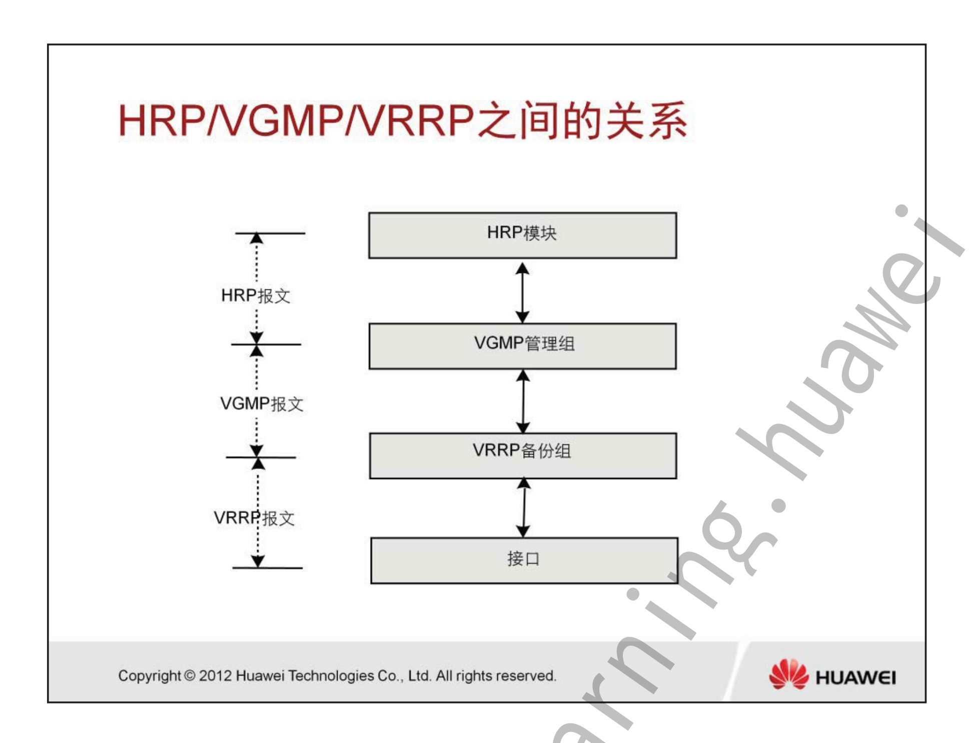
HRP模块提供了一个基础的数据备份机制和传输功能。各个应用模块收集本模块需要备份的数据，提交给HRP模块，HRP模块负责将数据发送到对端防火墙的对应模块，应用模块需要再将HRP模块提交上来的数据进行解析，并加入到防火墙的动态运行数据池中。

备份内容

- 要备份的连接状态数据包括TCP/UDP的会话表、ServerMap表项、动态黑名单、NO-PAT表项、ARP表项等。当备防火墙上没有这些数据时，切换到备防火墙上的流量可能会被防火墙阻拦，从而造成连接中断。

备份方向

- 只要防火墙上有状态为主的VGMP管理组，则它就会向对端备份。当两台防火墙上都有状态为主的管理组时，则连接状态数据会相互备份（但会保证不会重复备份）。在某些复杂的双机热备份组网中，同一台防火墙上可能配置多个VGMP管理组，且这些管理组的状态有可能不一致，这样导致主设备和备设备的界限变得模糊。



当VGMP管理组状态变化时，系统将通知HRP状态和配置主/从设备的状态发生相应的变化，从而确保两台防火墙之间配置命令和会话状态信息得到及时备份。同时，VGMP管理组状态也要受HRP状态影响，即VGMP会根据HRP状态切换的结果来调整优先级，并进行VGMP状态切换。

VGMP管理组报文和HRP模块的报文，都是通过主VRRP的接口进行传输，当VRRP接口接到的报文为VGMP管理组报文时，就交与VGMP模块进行处理；当接到的报文是VGMP的数据报文时，则通过VGMP模块与HRP模块的接口转到HRP模块进行处理。

当VRRP备份组状态变化时，由VGMP管理组来决定是否发生VGMP管理组的状态变化，并继而决定是否引起HRP和配置主/从设备的状态变化。



目 录



防火墙双机热备份组网方式

USG的双机热备份，可以工作在路由模式和混合模式两种模式下：

- 路由模式是指USG的业务端口和HRP备份通道接口均工作在路由模式下。
- 混合模式是指USG的业务端口工作在透明模式下，而HRP备份通道接口工作在路由模式下。

路由模式和混合模式都包含两种组网方式：

- 主备组网方式
- 负载分担组网方式

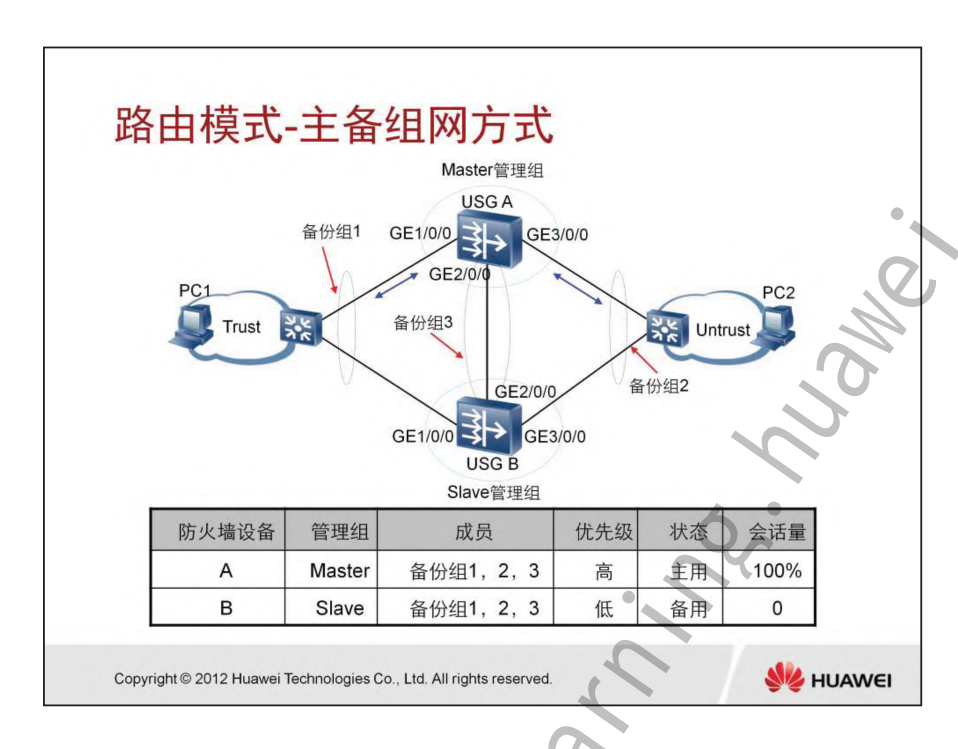


路由模式下的双机热备份

- 路由模式下的双机热备份主要是指USG通过路由模式的接口接收和发送业务报文、VRRP报文、VGMP报文和HRP报文。
- 为了提高可靠性，建议用户配置专用HRP备份通道，利用此通道传送VRRP报文、VGMP报文和HRP报文。

混合模式下的双机热备份

- 混合模式下的双机热备份主要是指USG通过透明模式的接口接收和发送业务报文，用以完成网络应用；通过路由模式的接口传送VRRP、VGMP和HRP报文，用以维护防火墙的主备关系。
- 混合模式下的双机热备份除能够提供透明模式的无缝接入，对外提供二层交换机（透明模式下的防火墙）业务外，还能保证当主用防火墙发生故障后，流量能够转换到备用防火墙上，保证业务的连续性。



防火墙设备	管理组	成员	优先级	状态	会话量
A	Master	备份组1，2，3	高	主用	100%
B	Slave	备份组1，2，3	低	备用	0

USG作为安全设备被部署在业务节点上。其中上下行设备均是交换机，USG A、USG B分别充当主用设备和备用设备，且均工作在路由模式下。

网络规划如下：

- 需要保护的网段地址为10.100.10.0/24，与防火墙的GigabitEthernet 1/0/0接口相连，部署在Trust区域。
- 外部网络与防火墙的GigabitEthernet 3/0/0接口相连，部署在Untrust区域。
- 两台防火墙的HRP备份通道接口GigabitEthernet 2/0/0部署在DMZ区域。
- 两台防火墙分别通过交换机连接各个安全区域。

其中，各安全区域对应的备份组虚拟IP地址如下：

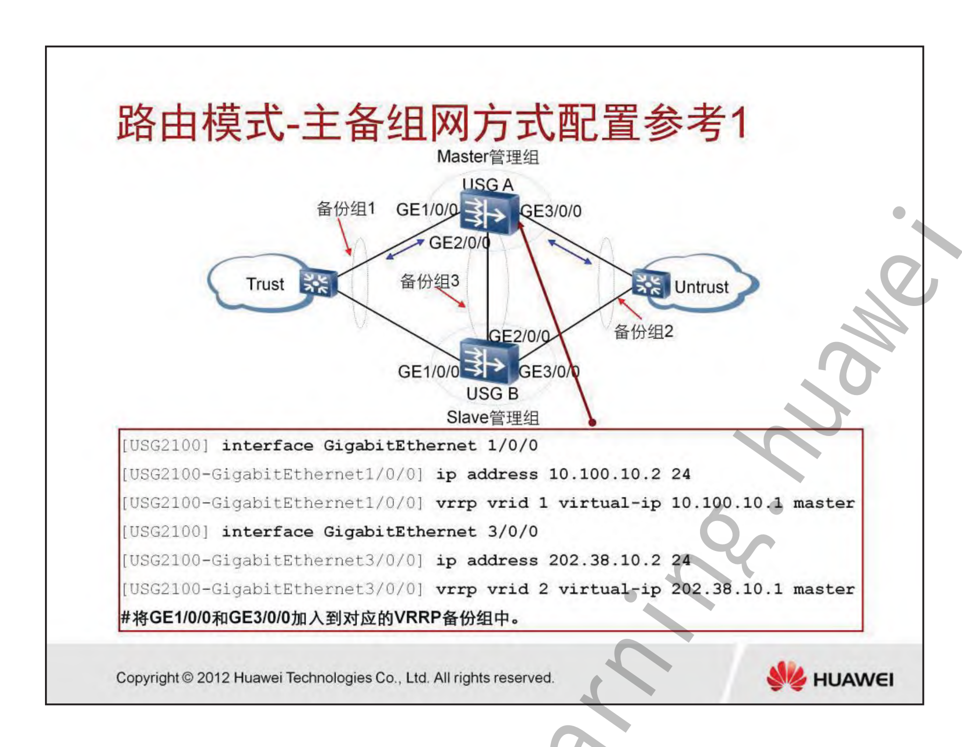
- Trust区域对应的备份组虚拟IP地址为10.100.10.1。
- Untrust区域对应的备份组虚拟IP地址为202.38.10.1。
- DMZ区域对应的备份组虚拟IP地址为10.100.20.1。

USG A

- GE1/0/0：10.100.10.2/24；GE2/0/0：10.100.20.2/24；GE3/0/0：202.38.10.2/24

USG B

- GE1/0/0：10.100.10.3/24；GE2/0/0：10.100.20.3/24；GE3/0/0：202.38.10.3/24

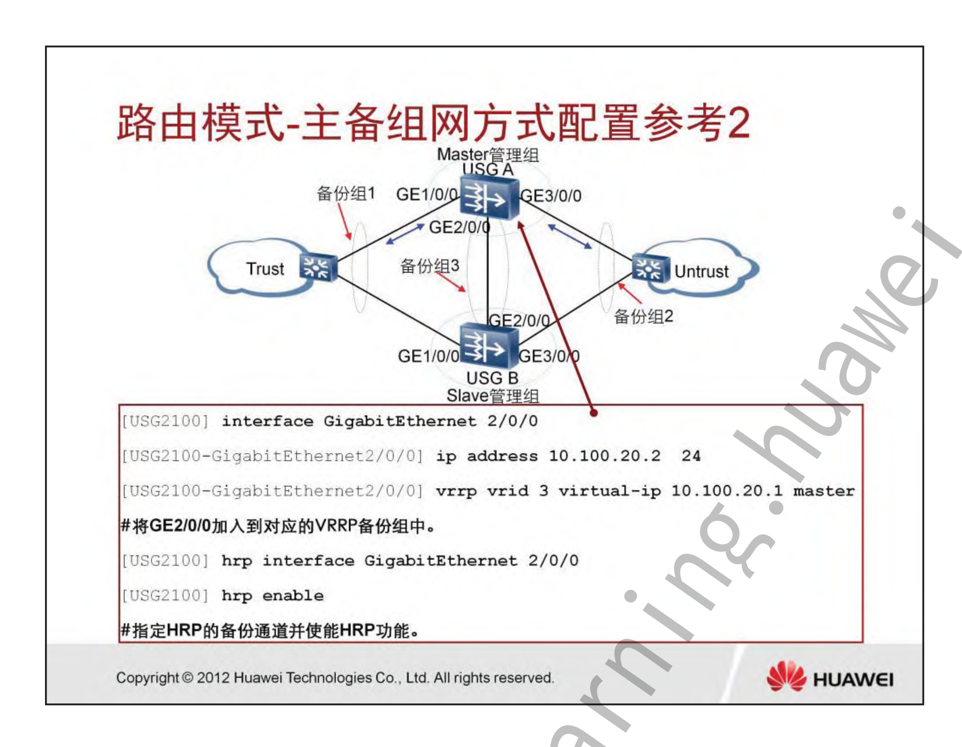


配置GigabitEthernet 1/0/0加入Trust区域。

- [USG2100] firewall zone trust
- [USG2100-zone-trust] add interface GigabitEthernet 1/0/0
- [USG2100-zone-trust] quit

配置GigabitEthernet 3/0/0加入Untrust区域。

- [USG2100] firewall zone untrust
- [USG2100-zone-untrust] add interface GigabitEthernet 3/0/0
- [USG2100-zone-untrust] quit

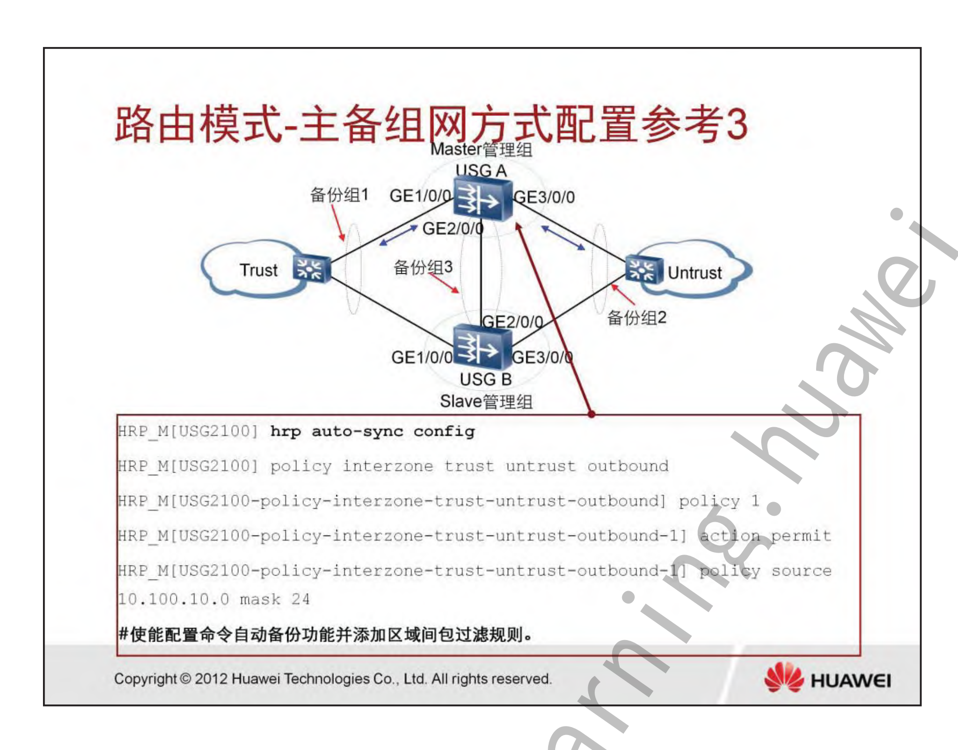


配置GigabitEthernet 2/0/0加入DMZ区域。

- [USG2100] firewall zone dmz
- [USG2100-zone-dmz] add interface GigabitEthernet 2/0/0
- [USG2100-zone-dmz] quit

配置USG B

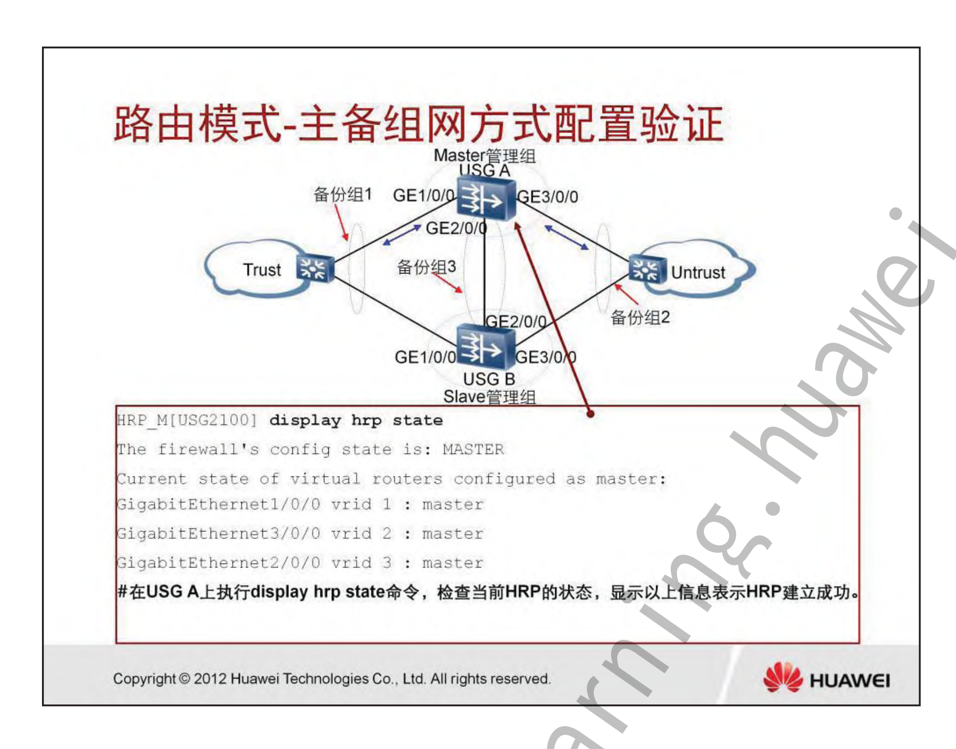
- USG B和上述USG A的配置基本相同，不同之处在于:
  - USG B各接口的IP地址与USG A各接口的IP地址不相同。
  - USG B的VRRP指定的管理组应该设为Slave。



当USG A和USG B都启动HRP功能并添加心跳线上区域间包过滤规则后，在USG A上开启配置命令的自动备份，这样在USG A上配置的ACL以及域间包过滤规则都将自动备份到USG B，不需要再在USG B上单独配置。

#添加区域间包过滤规则，使VRRP报文可以在两台防火墙之间的心跳线上交互。

- HRP_M[USG] policy interzone trust untrust outbound
- HRP_M[USG2100-policy-interzone-trust-untrust-outbound] policy 1
- HRP_M[USG2100-policy-interzone-trust-untrust-outbound-1] action permit
- HRP_M[USG2100-policy-interzone-trust-untrust-outbound-1] policy source 10.100.10.0 mask 24

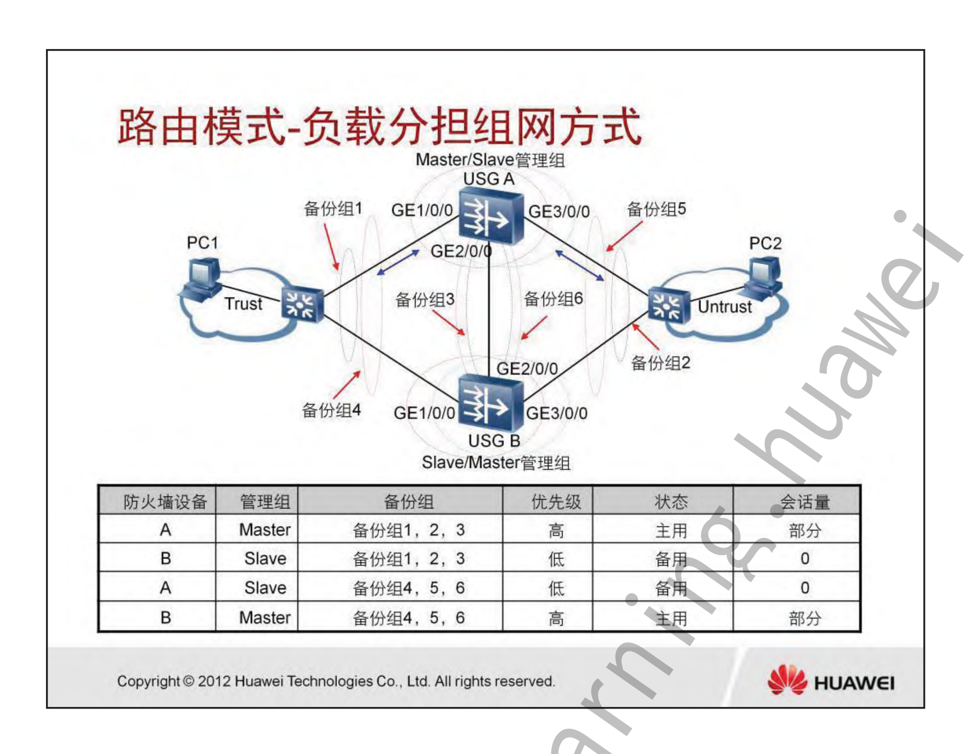


在USG A上执行display vrrp命令，检查VRRP备份组内接口的状态信息，显示以下信息表示VRRP备份组建立成功。

HRP_M[USG2100] display vrrp

- GigabitEthernet1/0/0 | Virtual Router 1
  - state : Master
  - Virtual IP : 10.100.10.1
  - Virtual MAC : 0000-5e00-0101
  - Primary IP : 10.100.10.2
  - PriorityRun : 100
  - PriorityConfig : 100
  - MasterPriority : 100
  - Preempt : YES
  - Delay Time : 0
  - Timer : 1
  - Auth Type : NONE
  - Check TTL : YES

- GigabitEthernet3/0/0 | Virtual Router 2
  - state : Master
  - Virtual IP : 202.38.10.1
  - Virtual MAC : 0000-5e00-0102
  - Primary IP : 202.38.10.2
  - PriorityRun : 100
  - PriorityConfig : 100
  - MasterPriority : 100
  - Preempt : YES
  - Delay Time : 0
  - Timer : 1
  - Auth Type : NONE
  - Check TTL : YES
- GigabitEthernet2/0/0 | Virtual Router 3
  - state : Master
  - Virtual IP : 10.100.20.1
  - Virtual MAC : 0000-5e00-0103
  - Primary IP : 10.100.20.2
  - PriorityRun : 100
  - PriorityConfig : 100
  - MasterPriority : 100
  - Preempt : YES
  - Delay Time : 0
  - Timer : 1
  - Auth Type : NONE
  - Check TTL : YES
  - hrp auto-sync config： 这条命令是启动配置命令的自动备份。此时通过display current 可以看到在防火墙A上配置的ACL同步到防火墙B上。



防火墙设备	管理组	备份组	优先级	状态	会话量
A	Master	备份组1，2，3	高	主用	部分
B	Slave	备份组1，2，3	低	备用	0
A	Slave	备份组4，5，6	低	备用	0
B	Master	备份组4，5，6	高	主用	部分

USG作为安全设备被部署在业务节点上。其中上下行设备均是交换机，USG A和USG B使用负载分担方式组网，且均工作在路由模式下。

网络规划如下：

- 需要保护的网段地址为10.100.10.0/24，与防火墙的GigabitEthernet 1/0/0接口相连，部署在Trust区域。
- 外部网络与防火墙的GigabitEthernet 3/0/0接口相连，部署在Untrust区域。
- 两台防火墙的HRP备份通道接口GigabitEthernet 2/0/0部署在DMZ区域。
- 两台防火墙分别通过交换机连接各个安全区域。

其中，各安全区域对应的备份组虚拟IP地址如下：

- Trust区域对应的备份组虚拟IP地址为：
  - 备份组1：10.100.10.1；备份组4：10.100.10.2
- Untrust区域对应的备份组虚拟IP地址为：
  - 备份组2：202.38.10.1；备份组5：202.38.20.2

DMZ区域对应的备份组虚拟IP地址为：

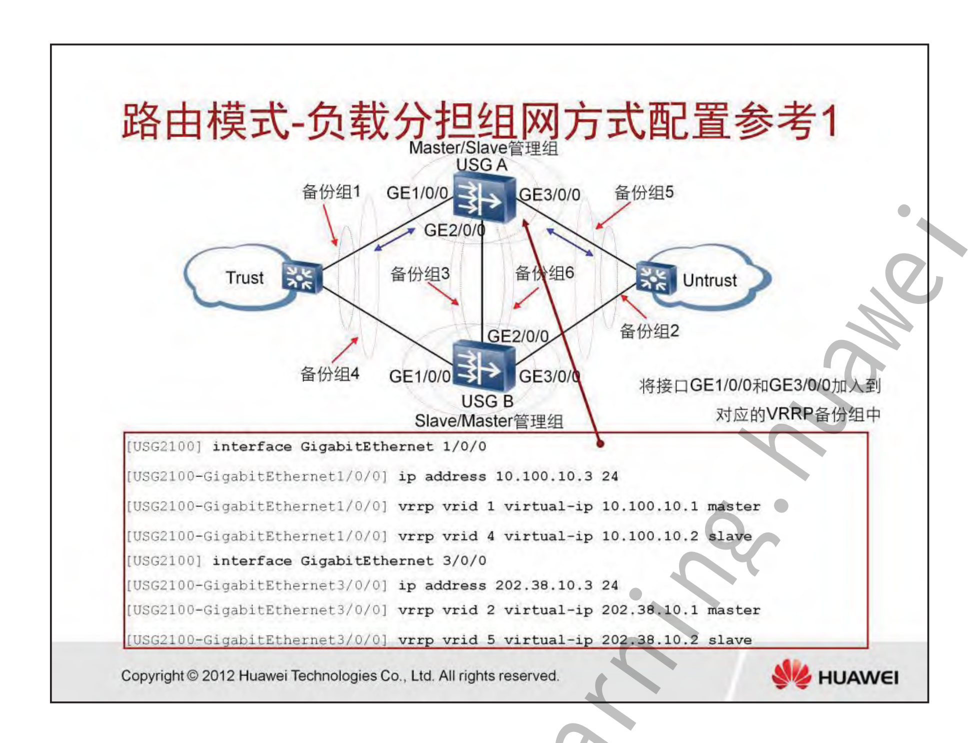
- 备份组3：10.100.20.1；备份组6：10.100.20.2

USG A

- GE1/0/0：10.100.10.3/24；GE2/0/0：10.100.20.3/24；GE3/0/0：202.38.10.3/24

USG B

- GE1/0/0：10.100.10.4/24；GE2/0/0：10.100.20.4/24；GE3/0/0：202.38.10.4/24

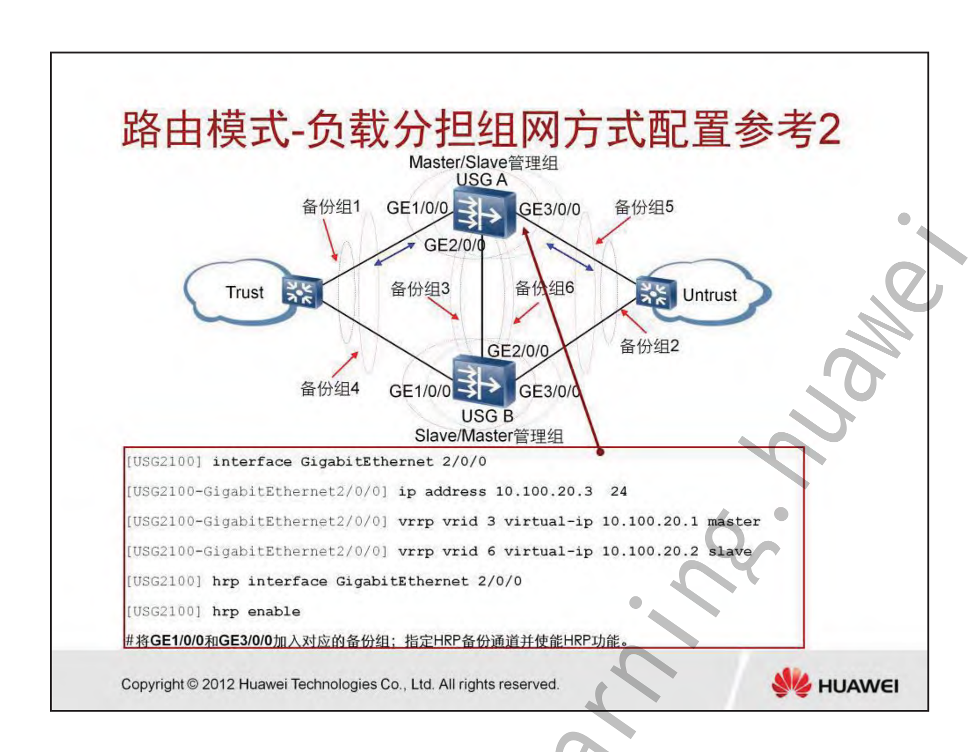


配置GigabitEthernet 1/0/0加入Trust区域。

- [USG2100] firewall zone trust
- [USG2100-zone-trust] add interface GigabitEthernet 1/0/0
- [USG2100-zone-trust] quit

配置GigabitEthernet 3/0/0加入Untrust区域。

- [USG2100] firewall zone untrust
- [USG2100-zone-untrust] add interface GigabitEthernet 3/0/0
- [USG2100-zone-untrust] quit

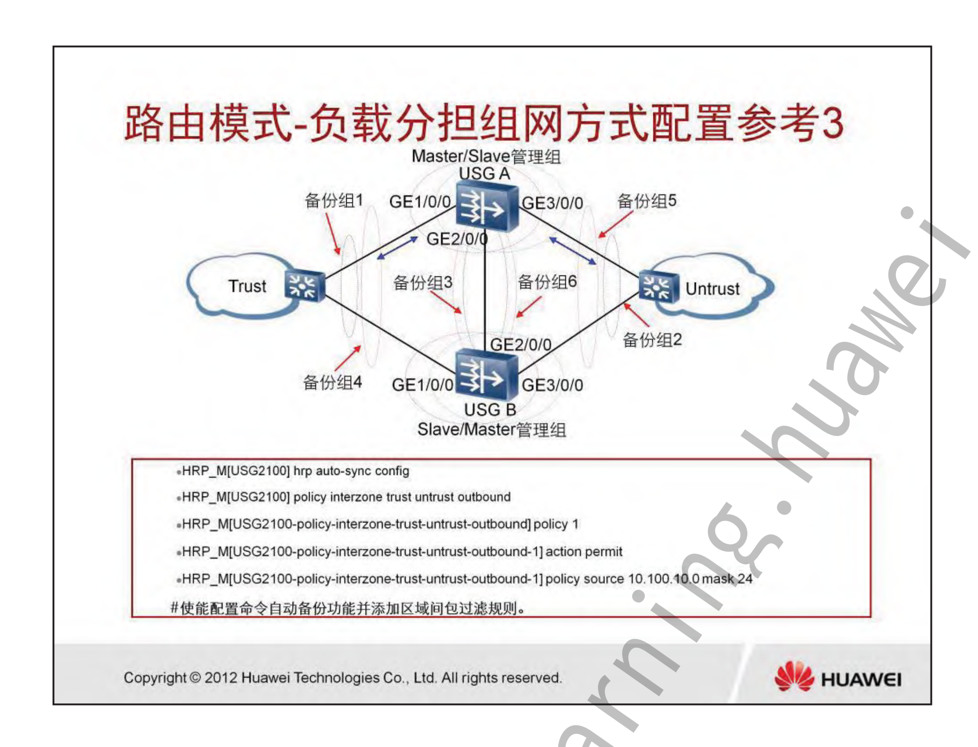


配置GigabitEthernet 2/0/0加入DMZ区域。

- [USG2100] firewall zone dmz
- [USG2100-zone-dmz] add interface GigabitEthernet 2/0/0
- [USG2100-zone-dmz] quit

配置USG B

- USG B和上述USG A的配置基本相同，不同之处在于:
  - USG B各接口的IP地址与USG A各接口的IP地址不相同。
  - 在USG A上被加入到Master管理组的VRRP备份组，在USG B上应该加入到Slave管理组中；在USG A上被加入到Master管理组的VRRP备份组，在USG B上应该加入到Slave管理组中。



当USG A和USG B都启动HRP功能并添加心跳线上区域间包过滤规则后，在USG A上开启配置命令的自动备份，这样在USG A上配置的ACL以及域间包过滤规则都将自动备份到USG B，不需要再在USG B上单独配置。

#添加区域间包过滤规则，使VRRP报文可以在两台防火墙之间的心跳线上交互。

- HRP_M[USG2100] policy interzone trust untrust outbound
- HRP_M[USG2100-policy-interzone-trust-untrust-outbound-1] policy 1
- HRP_M[USG2100-policy-interzone-trust-untrust-outbound-1] action permit
- HRP_M[USG2100-policy-interzone-trust-untrust-outbound-1] policy source 10.100.10.0 mask 24

路由模式-负载分担组网方式配置验证

```
HRP_M[USG2100] dis hrp state
The firewall's config state is: MASTER
 Current state of virtual routers configured as master:
            GigabitEthernet3/0/0    vrid   2 : master
            GigabitEthernet2/0/0    vrid   3 : master
            GigabitEthernet1/0/0    vrid   1 : master
 Current state of virtual routers configured as slave:
            GigabitEthernet3/0/0    vrid   5 : slave
            GigabitEthernet2/0/0    vrid   6 : slave
            GigabitEthernet1/0/0    vrid   4 : slave
```

#在USG A上执行display hrp state命令，检查当前HRP的状态。从以上显示信息可以看出，在USG A上，VRRP备份组1、2、3属于Master管理组；VRRP备份组4、5、6属于Slave管理组。



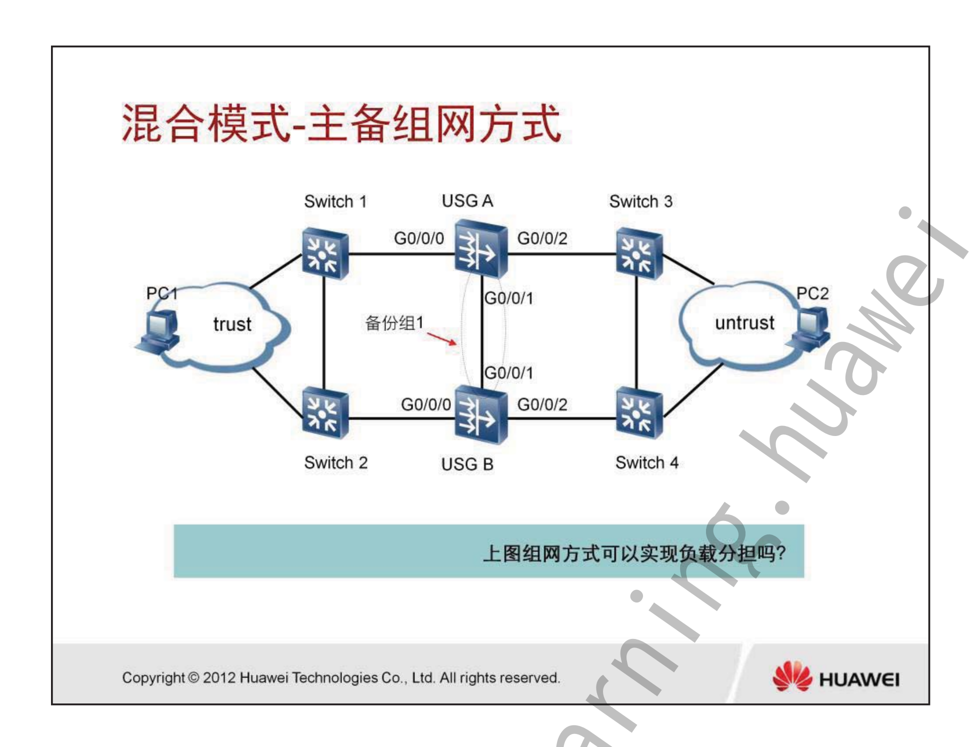
在USG A上执行display vrrp命令，检查VRRP备份组内接口的状态信息，显示以下信息表示VRRP备份组建立成功。

HRP_M[USG2100]dis vrrp

- GigabitEthernet3/0/0 | Virtual Router 2
- state : Master
- Virtual IP : 202.38.10.1
- PriorityRun : 100
- PriorityConfig : 100
- MasterPriority : 100
- Preempt : YES   Delay Time : 0
- Timer : 1
- Auth Type : NONE
- Check TTL : YES
- GigabitEthernet3/0/0 | Virtual Router 5
- state : Backup

- Virtual IP : 202.38.10.2
- PriorityRun : 100
- PriorityConfig : 100
- MasterPriority : 100
- Preempt : YES Delay Time : 0
- Timer : 1
- Auth Type : NONE
- Check TTL : YES
- GigabitEthernet2/0/0 | Virtual Router 3
- state : Master
- Virtual IP : 10.100.20.1
- PriorityRun : 100
- PriorityConfig : 100
- MasterPriority : 100
- Preempt : YES Delay Time : 0
- Timer : 1
- Auth Type : NONE
- Check TTL : YES
- GigabitEthernet2/0/0 | Virtual Router 6
- state : Backup
- Virtual IP : 10.100.20.2
- PriorityRun : 100
- PriorityConfig : 100
- MasterPriority : 100
- Preempt : YES Delay Time : 0
- Timer : 1
- Auth Type : NONE
- Check TTL : YES

- GigabitEthernet1/0/0 | Virtual Router 1
- state : Master
- Virtual IP : 10.100.10.1
- PriorityRun : 100
- PriorityConfig : 100
- MasterPriority : 100
- Preempt : YES   Delay Time : 0
- Timer : 1
- Auth Type : NONE
- Check TTL : YES
- GigabitEthernet1/0/0 | Virtual Router 4
- state : Backup
- Virtual IP : 10.100.10.2
- PriorityRun : 100
- PriorityConfig : 100
- MasterPriority : 100
- Preempt : YES   Delay Time : 0
- Timer : 1
- Auth Type : NONE
- Check TTL : YES



防火墙上下行设备是二层交换模块，并且是一种主备备份方式的组网，其中Switch-1、Swithc-3为主用设备；

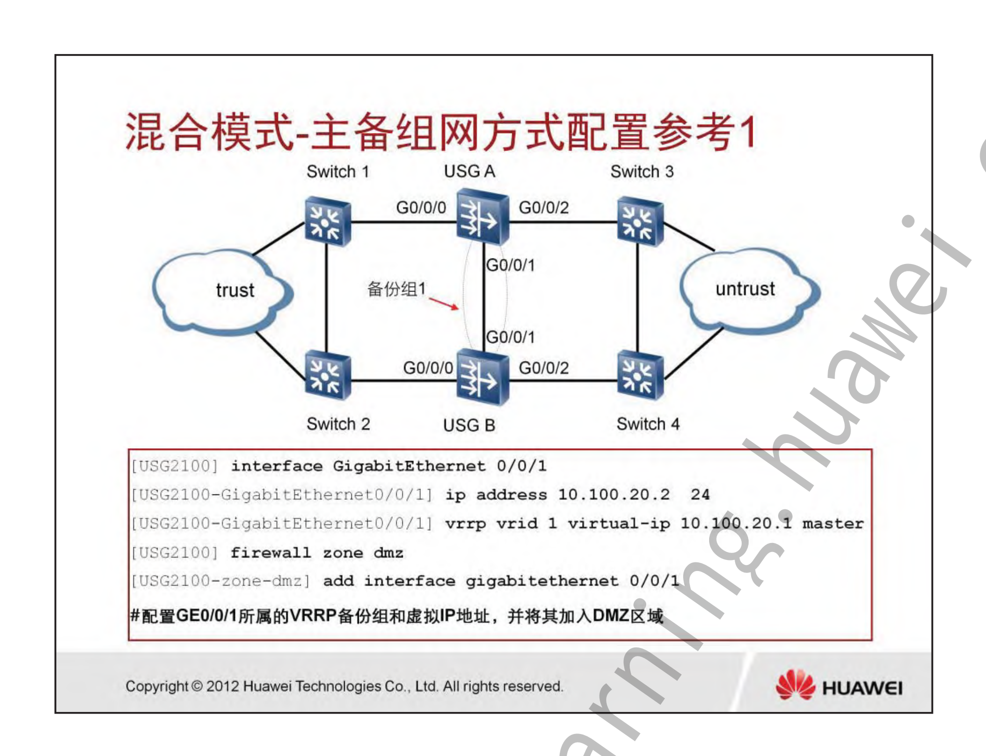
防火墙上下行业务端口工作在透明模式，心跳接口工作在路由模式；

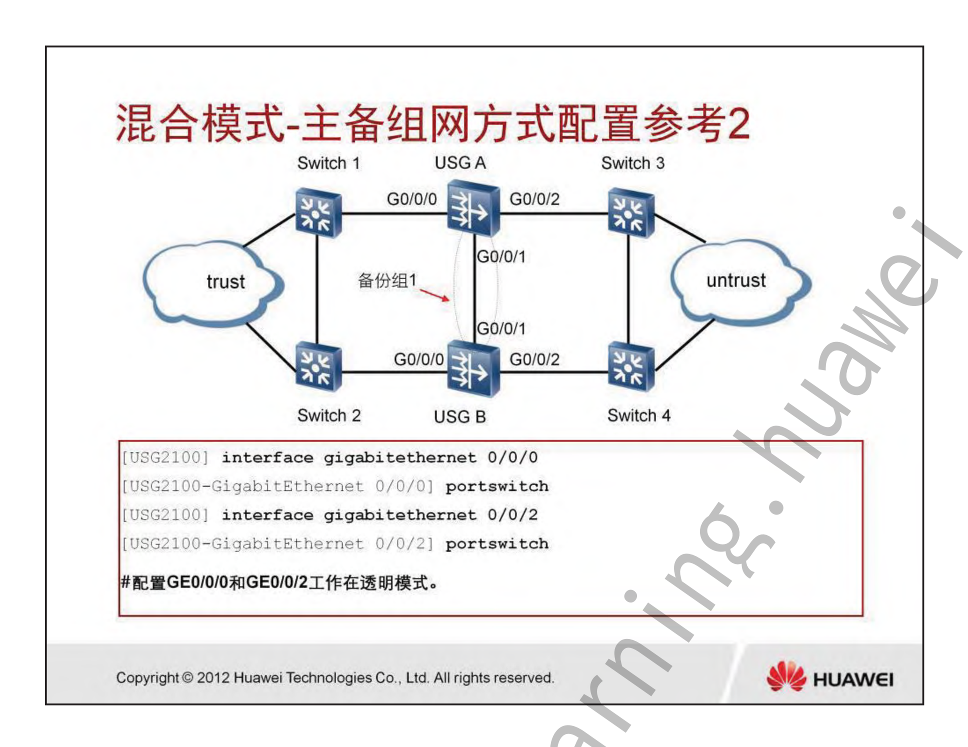
为了提高网络的可靠性，要求采用两台防火墙形成双机热备份；

防火墙USG A的G0/0/0、G0/0/1、G0/0/2接口分别位于trust、dmz和untrust区域，心跳接口GE0/0/1的IP地址为10.100.20.2/24，备份组虚拟IP地址为10.100.20.1/24；

防火墙USG B的G0/0/0、G0/0/1、G0/0/2接口分别位于trust、dmz和untrust区域，心跳接口GE0/0/1的IP地址为10.100.20.3/24，备份组虚拟IP地址为10.100.20.1/24。

注意：防火墙上下行端口工作在透明模式，无法配置IP地址，故PC1和PC2的IP地址应属于相同网段。


混合模式-主备组网方式配置参考1
Switch 1
USG A
Switch 3
G0/0/0
G0/0/2
G0/0/1
trust
备份组1
untrust
G0/0/1
G0/0/0
G0/0/2
Switch 2
USG B
Switch 4
[USG2100] interface GigabitEthernet 0/0/1
[USG2100-GigabitEthernet0/0/1] ip address 10.100.20.2 24
[USG2100-GigabitEthernet0/0/1] vrrp vrid 1 virtual-ip 10.100.20.1 master
[USG2100] firewall zone dmz
[USG2100-zone-dmz] add interface gigabitethernet 0/0/1
#配置GE0/0/1所属的VRRP备份组和虚拟IP地址，并将其加入DMZ区域
Copyright © 2012 Huawei Technologies Co., Ltd. All rights reserved.
HUAWEI



配置接口的工作模式之前，需要首先将全局的工作模式从默认的路由模式更改为混合模式。注意：更改模式后需要重启防火墙设备。

进入Trust区域视图。

[USG2100] firewall zone trust

配置GigabitEthernet 0/0/0加入Trust区域。

[USG2100-zone-trust] add interface gigabitethernet 0/0/0

进入Untrust区域视图。

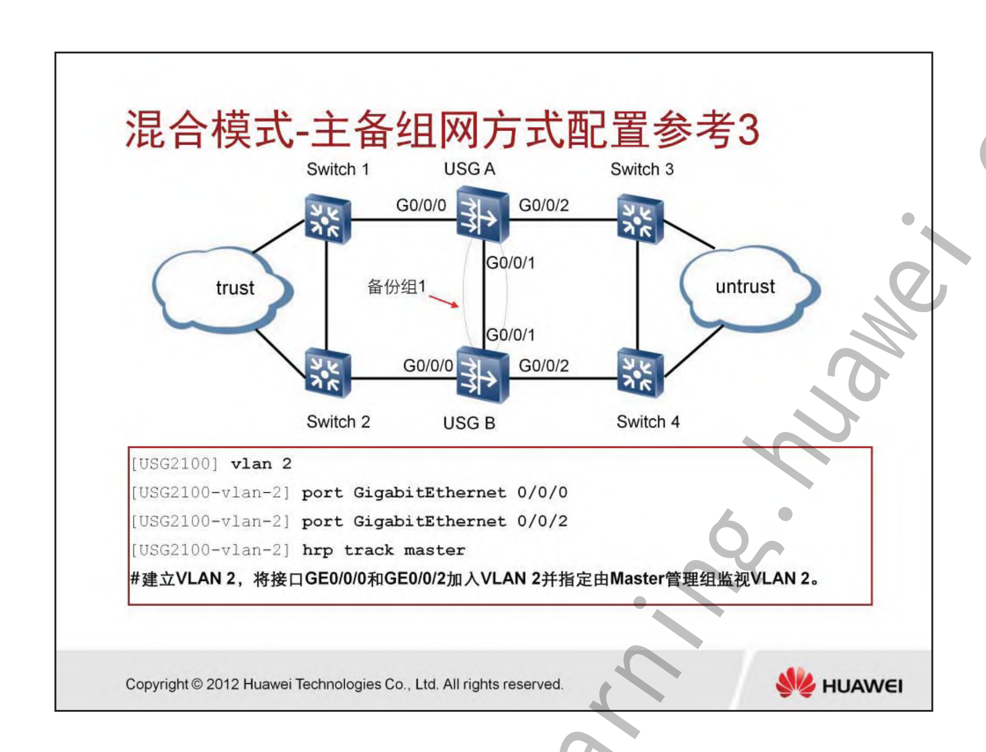
[USG2100] firewall zone untrust

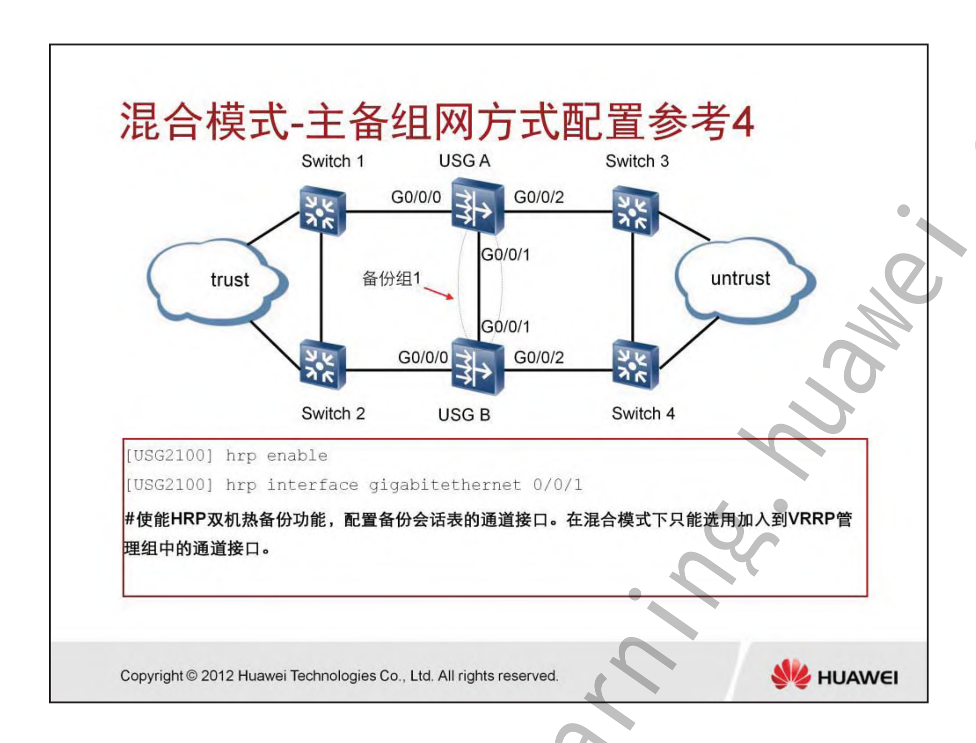
配置GigabitEthernet 0/0/2加入Untrust区域。

[USG2100-zone-untrust] add interface gigabitethernet 0/0/2

配置统一安全网关域间缺省规则为允许所有报文通过（可以根据实际需要使用其他配置策略）。

[USG2100] firewall packet-filter default permit all


混合模式-主备组网方式配置参考3
Switch 1
USG A
Switch 3
G0/0/0
G0/0/2
G0/0/1
trust
备份组1
untrust
G0/0/1
G0/0/0
G0/0/2
Switch 2
USG B
Switch 4
[USG2100] vlan 2
[USG2100-vlan-2] port GigabitEthernet 0/0/0
[USG2100-vlan-2] port GigabitEthernet 0/0/2
[USG2100-vlan-2] hrp track master
#建立VLAN 2，将接口GE0/0/0和GE0/0/2加入VLAN 2并指定由Master管理组监视VLAN 2。
Copyright © 2012 Huawei Technologies Co., Ltd. All rights reserved.
HUAWEI



USG B和USG A的配置绝大多数相同，差别如下：

- USG B上的接口IP地址与USG A不同。
- USG B上指定由Slave管理组监视VLAN。
- 注意：HRP备份通道接口不能为二层交换接口或VlanIf接口。

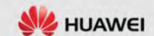
混合模式-主备组网方式配置验证

```
HRP_M[USG2100] display hrp state
The firewall's config state is: MASTER
Current state of virtual routers configured as master:
        GigabitEthernet0/0/1    vrid   1 : master
```

#在USG A上执行display hrp state命令，检查当前HRP的状态。从以上显示信息可以看出，在USG A上，VRRP备份组1属于Master管理组。

```
HRP_S[USG2100] display hrp state
The firewall's config state is: SLAVE
Current state of virtual routers configured as slave:
        GigabitEthernet0/0/1    vrid   1 : slave
```

#在USG B上执行display hrp state命令，检查当前HRP的状态。从以上显示信息可以看出，在USG B上，VRRP备份组1属于Slave管理组。



在USG A上执行display vrrp命令，检查VRRP备份组内接口的状态信息，显示以下信息表示VRRP备份组建立成功。

HRP_M[USG2100]dis vrrp

- GigabitEthernet0/0/1 | Virtual Router 1

  state : Master

  Virtual IP : 10.100.20.1

  PriorityRun : 100

  PriorityConfig : 100

  MasterPriority : 100

  Preempt : YES   Delay Time : 0

  Timer : 1

  Auth Type : NONE

  Check TTL : YES

在USG B上执行display vrrp命令，检查VRRP备份组内接口的状态信息，显示以下信息表示VRRP备份组建立成功。

HRP_S[USG2100]dis vrrp

- GigabitEthernet0/0/1 | Virtual Router 1

state : Backup

Virtual IP : 10.100.20.1

PriorityRun : 100

PriorityConfig : 100

MasterPriority : 100

Preempt : YES　Delay Time : 0

Timer : 1

Auth Type : NONE

Check TTL : YES

问 题

1、VGMP管理组的主要功能有哪些?

2、请列举常见的防火墙组网方式。



答案:

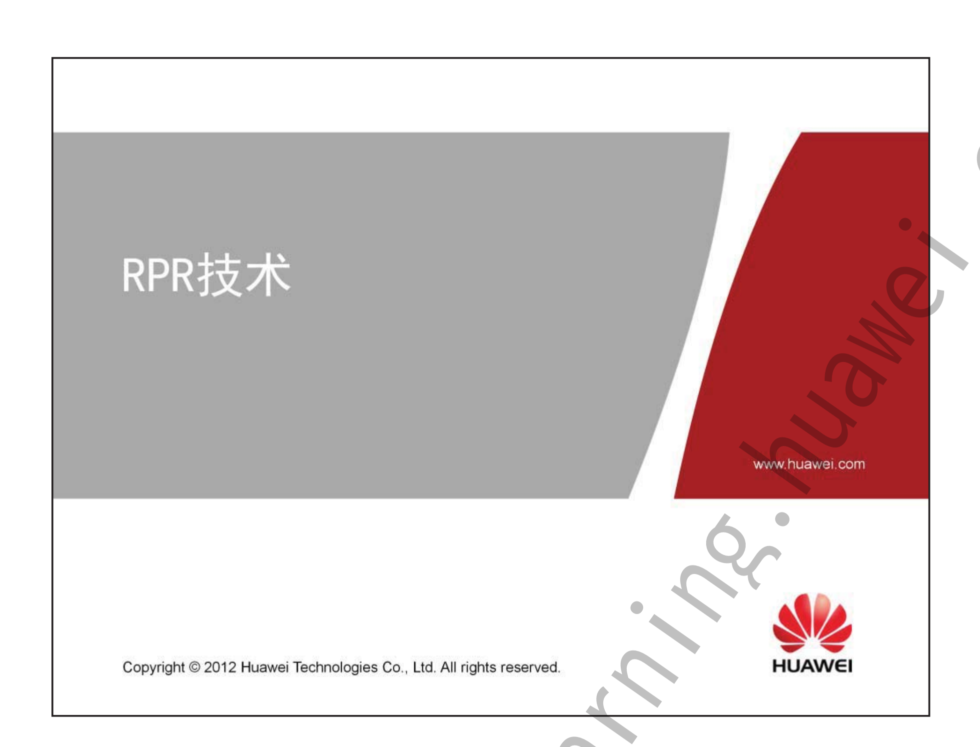
- 1、状态一致性管理；抢占管理；通道管理。
- 2、路由模式-主备组网方式；路由模式-负载分担组网方式；混合模式-主备组网方式；混合模式-负载分担方式。

谢谢

www.huawei.com

Module 2

可靠性


RPR技术
www.huawei.com
Copyright © 2012 Huawei Technologies Co., Ltd. All rights reserved.
HUAWEI



前 言

本课程介绍弹性分组数据环（RPR）技术的基本原理与配置。

RPR技术综合了SDH/SONET和以太网以及其它一些环网技术的优点，集IP的智能化、以太网的经济性和光纤环网的高带宽、高可靠性于一体。提供一种更优的城域网解决方案。


Page1


SDH/SONET（Synchronous Digital Hierarchy/Synchronous Optical Network）环网的优点是高可靠性，满足用户的通信要求；能够提供保护和快速恢复机制；但是其点到点、电路交换的设计目标也为它带来了诸多缺点：1）带宽在节点间点到点的链路中固定分配并保留；2）带宽不能根据网络中流量的实际情况而改变，不利于带宽的高效利用；3）广播和组播报文将分成多个单播完成，浪费带宽；4）通常为实现保护机制，50% 的带宽将保留，未能提供灵活的选择机制。

以太网技术具有成本低、简洁、易扩展、以及便于IP 包的传输和处理等特点，但它在规模、端到端业务建立、质量保证、可靠性等方面还存在不少需要克服的难题。



培训目标

学完本课程后，您应该能：

- 理解RPR技术的基本概念
- 掌握RPR技术的基本原理

Page2
HUAWEI


目 录
环网技术概述
RPR环技术概述
RPR环技术特点
RPR环的基本概念
RPR环的数据操作
RPR环的保护倒换
RPR环的公平算法
Copyright © 2012 Huawei Technologies Co., Ltd. All rights reserved.
Page3
HUAWEI

环网技术概述

环网技术就是将一些网络设备，通过环型拓扑结构连接到一起，实现相互通信的一种技术。

环网技术：

- Token Ring
- FDDI
- SDH/SONET
- POS/GE
- RPR 环

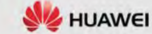
 Page4 HUAWEI



环网技术概述（续）

Token Ring〔令牌环〕是最早引入数据通信领域中的环网技术，是一个基于MAC层协议的单向环网，用于局域网，不具备故障自愈的保护功能。令牌环网是一种低速网络，一般在5类线缆上面传送。令牌环网的节点只有在获得令牌的情况下才能向环上发送数据，令牌逐点传送，每个节点能轮流拥有一定的令牌时间，节点需要等待令牌以传送数据。已有令牌的节点，如果没有数据需要传送，可以将令牌传递给下一个节点。数据包采用源节点剥离的方法，即数据包被送到环网上后，经过目的节点接收后，还会继续在环上转一圈，直到回到源节点才被剥离，显然，这种方式下，整个环上的某同一时刻，只能有一个节点可以传送数据。

 Page5



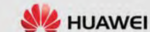
Token Ring〔令牌环〕是最早引入数据通信领域里面的环网技术，是一个基于MAC层协议的单向环网。其特点：

- 一种低速网络
- 节点只有获得令牌时才能向环上发送数据
- 数据包采用源节点剥离的方法
- 不具备故障自愈的保护功能

环网技术概述（续）

FDDI〔光纤分布式数据接口〕可以说是一种改进的Token Ring技术，也是利用令牌来传递对环网的控制权，所不同的是，他采用了双环结构，采用光纤作为传输介质，在性能和效率上都较令牌环网有很大提高。但是FDDI和Token Ring一样，不具备故障自愈的保护功能。因为也是采用源地址剥离技术，带宽得不到有效利用。FDDI网络目前在企业网和校园网中得到广泛的应用。

Page6



FDDI〔光纤分布式数据接口〕可以说是一种改进的Token Ring技术。其技术特点：

- 双环结构
- 采用光纤作为传输介质
- 利用令牌来传递对环网的控制权
- 数据包采用源节点剥离的方法
- 不具备故障自愈的保护功能



环网技术概述（续）

SDH/SONET〔数字同步系列〕是目前广泛应用在传输网络里面的一种环网技术，支持单环、多环，具有高可靠性，能提供故障自动保护倒换APS故障自愈机制。 SDH/SONET采用点到点电路交换的设计，环内带宽被静态分配为一条条静态固定带宽链路，使用时分复用，主要为语音服务。由于其点到点电路交换的设计，带来了很多缺点，如逻辑全连接时带宽浪费严重，带宽在节点间点到点的链路中固定分配并保留，带宽不能根据网络中流量的实际情况而改变，不利于带宽的高效利用，很难适应具有突发性特点的IP数据业务。广播和组播报文将分成多个单播传送，带宽浪费严重，而且对于APS特性，需要最高多达50%冗余带宽，未能提供灵活的选择机制。

 Page7



SDH/SONET〔数字同步系列〕是目前广泛应用在传输网络里面的一种环网技术 。其技术特点：

- 单环、多环都支持
- 采用点到点、电路交换
- 采用时分复用
- 主要为语音服务
- 高可靠性

环网技术概述（续）

GE/POS，严格来说，并不是一种环网技术，仅仅是近年来，网络上的一种比较流行的组网应用。将网络上面的N个节点通过N条链路首尾相连起来，整个环其实是由N个相互独立的点到点POS/PPP连接构成的，业务在节点间逐点三层转发实现相互通信，节点需要处理所有报文，严重影响环的吞吐量。由于仅仅是一个组网性的应用，没有一个针对环级别的带宽管理，某段的拥塞无法通知其它节点，而且没有二层的故障自愈能力。

 Page8 

POS环，严格来说，并不是一种环网技术，仅仅是近年来，网络上的一种组网应用。将网络上面的N个节点通过N条链路首尾相连起来，整个环其实是由N个相互独立的点到点POS/PPP连接构成的。其技术特点：

- 业务在节点间逐点三层转发实现相互通信
- 无带宽管理方法
- 不具备故障自愈的保护功能，倒换完全由路由协议控制。


目 录
环网技术概述
RPR环技术概述
RPR环技术特点
RPR环的基本概念
RPR环的数据操作
RPR环的保护倒换
RPR环的公平算法
Copyright © 2012 Huawei Technologies Co., Ltd. All rights reserved.
Page9
HUAWEI

RPR环技术概述

RPR技术综合了SDH/SONET和以太网以及其它一些环网技术的优点，研究并规范化一种环网拓扑上使用的MAC 层协议－RPR（弹性分组环），满足未来网络的要求。集IP的智能化、以太网的经济性和光纤环网的高带宽、高可靠性于一体。提供一种更优的城域网解决方案。

RPR的设计目标定义了一个闭合环路、点到点、基于MAC层的逻辑环状拓扑。对于物理层来说，RPR就是一组点到点的链路；而对于数据链路层来说，RPR就像是一个类似于Ethernet的广播介质网络。

Page10


RPR技术使得运营商在城域网内以低成本提供电信级的服务成为可能，在提供类似SDH级网络可靠性的同时降低了传送费用。RPR有别于传统MAC最吸引人的特点是具有电信级的可靠性，使其不仅仅只是局限于处理面向数据的业务传送需求，同时可以形成处理多业务传送的综合传输解决方案。

RPR是IP技术与光网络技术直接融合的产物，它源于客户对IP业务发展的需求，顺应最新的技术潮流，为IP城域网的建设带来了一套低成本、高品质的解决方案。

RPR（Resilient Packet Ring）协议是一个工作在OSI（Open Systems Interconnection）协议栈的数据链路层的介质访问控制协议，主要应用在城域网中。

RPR需要专用硬件支持 。




目 录
环网技术概述
RPR环技术概述
RPR环技术特点
RPR环的基本概念
RPR环的数据操作
RPR环的保护倒换
RPR环的公平算法
Copyright © 2012 Huawei Technologies Co., Ltd. All rights reserved.
Page11
HUAWEI

RPR环技术特点

物理层多样性

带宽高利用率

快速保护机制

公平的节点带宽分配

拓扑自动发现机制，支持即插即用

有效支持组播和广播

流量等级保证QoS，支持带宽预留的业务

 Page12 

物理层多样性:

- RPR（弹性分组环）技术是一种在环型结构上优化数据业务传送的新型MAC层协议，能够适应多种物理层(如SDH、以太网、DWDM等)

公平的带宽分配协议，提高带宽利用率:

- 环网的资源在节点之间是共享的，RPR提供一种整个环网级别的全局公平算法，以保证节点间公平享用带宽，并尽力提高带宽的最大利用率。公平算法能够动态地对网络流量进行调控，尽量避免网络拥塞，对于突发的大数据流量进行有效地调节，保证用户正常地使用网络。为实现这一目标，RPR 环网节点监测自身带宽资源的使用情况，同时在节点间提供显式的反压机制，该反馈信息通告发送源网络当前的可用能力，使之调整流量，最终实现全网的公平。

快速保护机制:

- 对于环上突发的故障，RPR协议可以迅速响应，进行保护，保证业务在50ms内恢复。 具有网络拓扑结构的自动发现和更新功能:
- 在网络拓扑变化时，每个节点通过接收RPR环上其它节点的MAC地址，自动建立和更新自己的拓扑图，使得网络初始化配



置变得极其简单，实现了即插即用，并可避免手工配置带来的错误，便于进行网络的运营维护 。

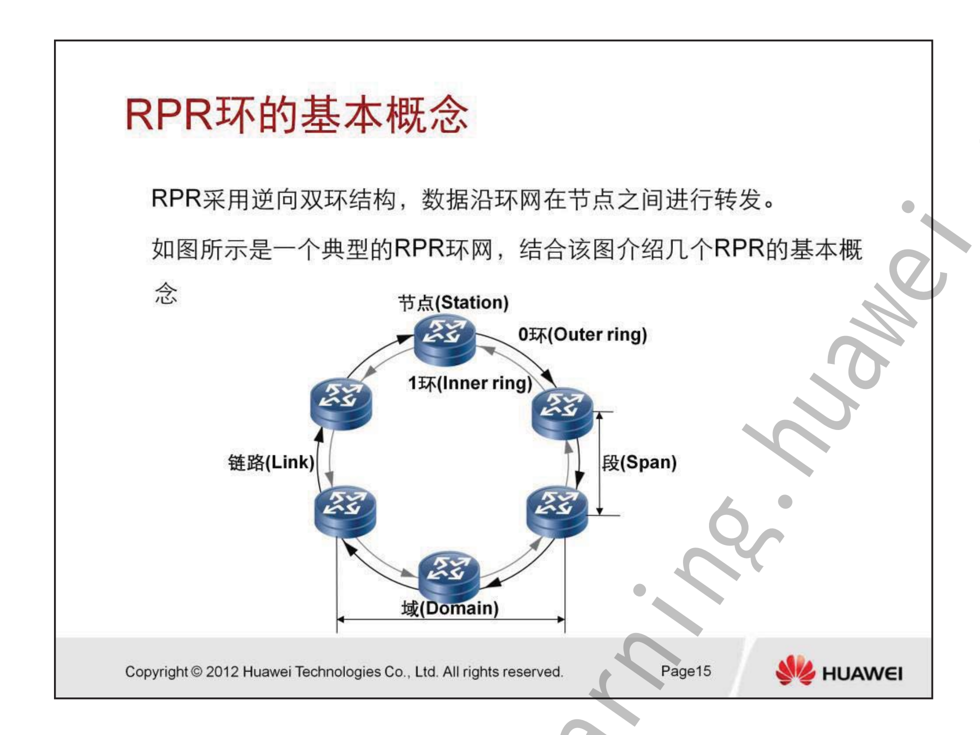
支持单播、组播和广播:

- 可将基于IEEE 802.3MAC地址的单播、组播和广播数据包映射到节点的RPR MAC地址，实现在RPR环路上根据节点的RPR MAC地址完成单播、组播和广播数据业务的传送。

QOS保证:

- RPR天然具备很好的QOS保证，比如50ms故障自愈能力、高带宽利用率、先进RPR-Fa公平算法等，可以提供高可靠性、大吞吐量、小迟延、低丢失率的业务保证能力。RPR支三种流量等级 。


目 录
环网技术概述
RPR环技术概述
RPR环技术特点
RPR环的基本概念
RPR环的数据操作
RPR环的保护倒换
RPR环的公平算法
Copyright © 2012 Huawei Technologies Co., Ltd. All rights reserved.
Page14
HUAWEI



每个RPR节点(station)都采用了一个以太网中用到的48位MAC地址作为地址标识，因此从RPR节点设备链路层来看，这两对收发的物理光接口只是一个链路层接口；从网络层来看，也只需要分配一个接口IP地址。

RPR环的基本概念（续）

- 0环（ringlet0）：RPR双环中，数据帧发送方向为顺时针的称为0环，也称“外环”。
- 1环（ringlet1）：RPR双环中，数据帧发送方向为逆时针的称为1环，也称“内环”。
- 节点（Station）：RPR环网上的设备，它负责接收和转发数据。
- 链路（Link）：连接相邻节点之间的一段传输通道。相邻节点之间由方向相反的两条链路连接。
- 段（Span）：RPR环网上两个相邻节点之间的链路，由方向相反的两条链路组成。

 Page16 

外环和内环都传送数据包和控制包，内环的控制包携带外环数据包的控制信息，反之亦然。

RPR环的基本概念（续）

域（Domain）：多个连续的段和其上的节点构成域。

节点的东向连接段：指和节点相邻的一个段，该段位于节点的1环入接口所在的一侧。

节点的西向连接段：指和节点相邻的一个段，该段位于节点的0环入接口所在的一侧。

边（Edge）：当段或和段相邻的节点出现故障时，段不能转发数据就成为边。

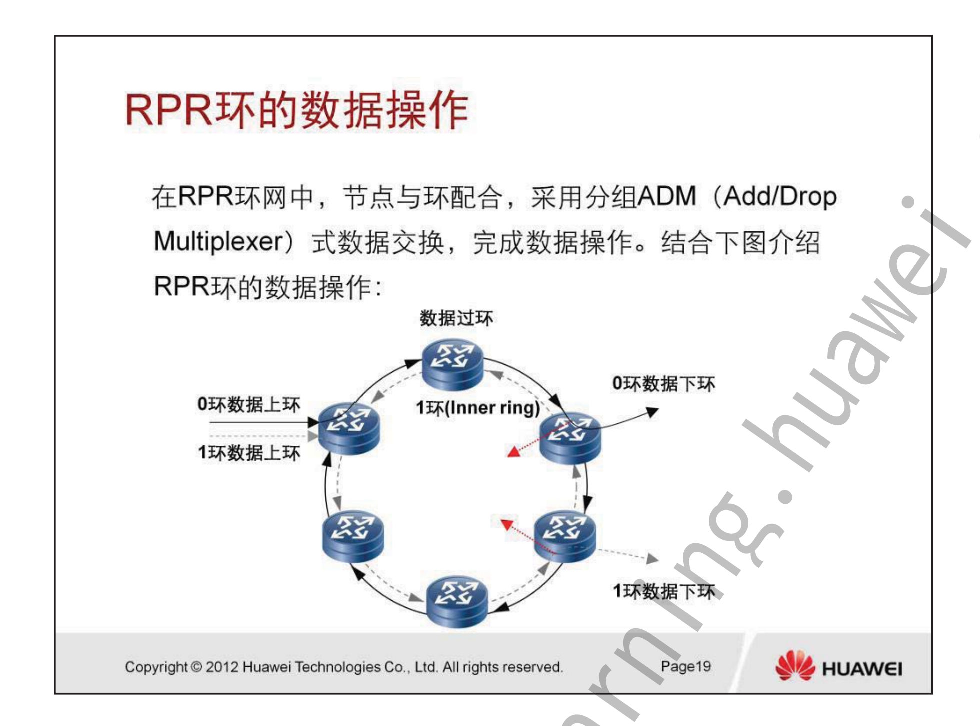
Page17



东向物理层的“发送口”与西向物理层的“接收口”通过MAC实体连接在一起，构成RPR的0环。

东向物理层的“接收口”与西向物理层的“发送口”相连，构成RPR的1环。


目 录
环网技术概述
RPR环技术概述
RPR环技术特点
RPR环的基本概念
RPR环的数据操作
RPR环的保护倒换
RPR环的公平算法
Copyright © 2012 Huawei Technologies Co., Ltd. All rights reserved.
Page18
HUAWEI



这里对于过环的数据操作，与SDH ADM设备的处理方式很相似，即过环数据流不需要设备上层处理，这样一来，设备处理性能大大提高。这种数据分组的ADM式交换体系很容易支撑各种高速链路接口。

RPR环的数据操作（续）

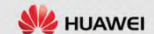
上环（insert）：节点设备把来自环网外的数据帧插入到RPR环网的数据流中。

下环（copy）：节点设备从RPR环网的数据流中接收数据帧，并将数据帧交给节点上层作相应处理。

过环（transit）：节点设备将途经本节点的数据流转发到下一个节点。

剥离（strip）：节点设备不再往下转发途经本节点的数据，即终止数据帧在RPR环网上的转发。

 Page20



RPR环作为一个共享的媒体介质，RPR 环上的每个结点对数据包的处理主要有四种操作：Insert、transit、copy和Strip，简化了结点间的操作。

Insert：数据包上环操作。RPR端口发送的数据包是经过L3层从其他端口转发而来，这些报文的下一条节点均是环上节点。当然，一些路由协议报文会直接有业务层生成，进入环网。

copy：接收从环上来的数据报，送Host/L3处理。

Strip：数据包下环操作。对于目的Mac为本节点的数据包，节点将执行剥离操作，并将数据送往L3处理。在L3，依据报文IP地址执行转发，转发到其它端口，或有本节点的业务层接收，如路由协议报文。

transit：只是经由本节点去往其它节点的数据包，本节点在L2执行快速转发，L2转发完全由硬件完成。极大的提高了RPR节点的L3吞吐量。而对于其它环技术，所有的包，无论是否由本节点接收的，都需要送到L3处理，不适合大流量网络。对于多播业务，节点在执行L2转发的同时，会将数据包送往L3处理。RPR节点特有Passthough工作模式，对所有报文简单执行transit操作。

Pass though 模式，在这种模式下，一个RPR节点相当于一环上的一个透明节点，对上下节点间的所有包进行完全转发，包括Usage Message包。在其它节点的拓扑信息里面没有该节点，但是原来的物理双环结构仍然保持。

在Shut Down或者上层失效的情况下，节点会自动进行Pass though模式，以保证环的持续可用性，带宽不受影响。而IPS自动保护发生时，环带宽其实减少了一半，因为环已经从正常的双环结构变为了单环结构。

RPR环的数据操作（续）

对于单播流量，在源节点处，采用上环操作，使数据承载到0环或1环中。目的节点执行数据下环和剥离操作。而中间节点只执行数据过环操作。值得注意的是，对于单播流量，RPR采取的是目的节点剥离的方式，报文一旦到达目的地，就不再在环上继续传送。这一点不同于传统环网技术所采用的源节点剥离。目的节点剥离能够有效提高带宽利用率，使得带宽的空间重用技术更高效。

对于组播和广播流量，由于有多个目的节点，在目的节点会同时执行过环和下环操作。相应的数据包在环网上只有一份拷贝。多播和广播是基于源剥离的，即目的节点将接收数据包并转发，而源节点则负责将多播包和广播包从环网上剥离。

 Page22 

上环数据该如何选择0环还是1环?

- 每个RPR 节点存储着一张选环表，表内记录着从本节点向RPR 环网上其它节点发送数据帧的路径，即从0环发送还是从1环发送。选环表中的表项分为动态选环表项和静态选环表项两种。
  - 动态表：动态选环表项是从本节点向RPR 环网上其它节点发送数据帧的最优路径。RPR 节点依靠拓扑自动发现功能自动维护动态选环表项。最优路径是指从本节点到目的节点距离最近（即跳数最少）的路径，当0环和1环上计算出的跳数相同时，选择0环作为最优路径。
  - 静态表：手动配置的选环表项称为静态表项。与动态选环表项不同，静态选环表项一旦配置就不会老化。通过合理地配置静态选环，可以更好的利用环网带宽，避免一个环流量极大，而另一个环流量较小。

RPR 环网对于组播、广播和未知单播数据帧的剔除有两种方式:



源节点剔除和TTL剔除。

- 源节点；剔除是指当数据帧沿环网传送一周再回到上环节点时，被上环节点从环网上剔除。
- TTL剔除：是指当数据帧的TTL字段值为0时，节点将该数据帧从环网上剔除。


目 录
环网技术概述
RPR环技术概述
RPR环技术特点
RPR环的基本概念
RPR环的数据操作
RPR环的保护倒换
RPR环的公平算法
Copyright © 2012 Huawei Technologies Co., Ltd. All rights reserved.
Page24
HUAWEI

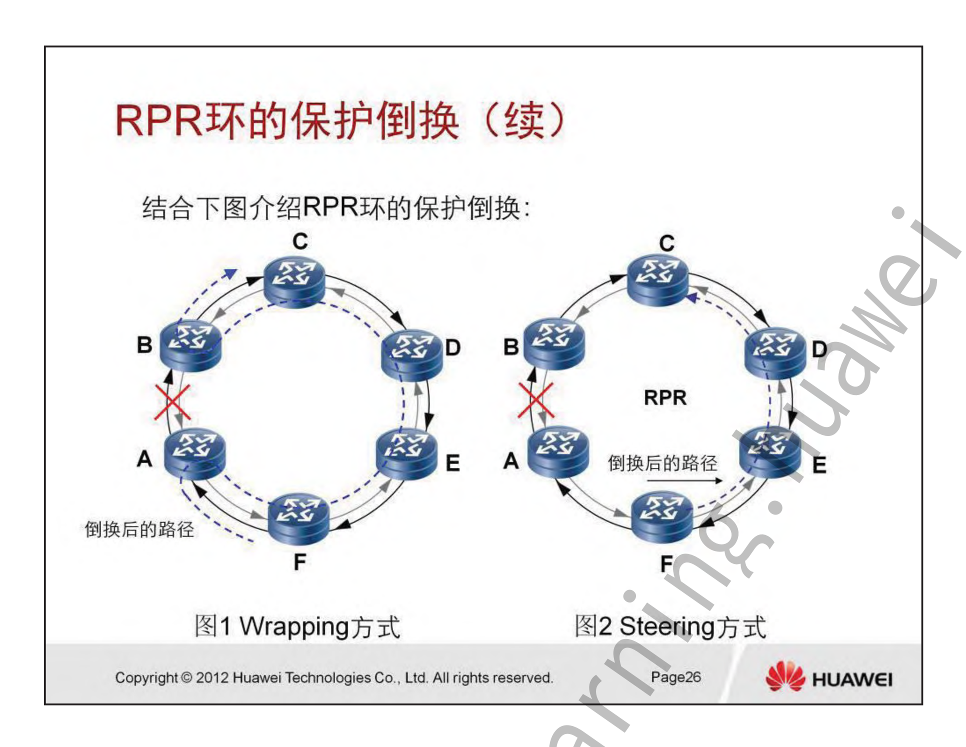
RPR环的保护倒换

当RPR环上节点失效或者链路发生故障的时候，RPR可以通过自动保护倒换保证环网的连通性，提供相当于SDH APS的低于50毫秒故障保护能力。RPR的自动保护倒换不需要像SDH一样的50%的冗余带宽开销。RPR可以对物理媒质的故障以及L2层的节点失效进行有效的保护。

 Page25



RPR采用了SDH的环形结构，同时也继承了一个特点，就是故障自愈能力非常强，能够实现50ms时间内的故障保护切换。



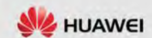
2种保护倒换方式

- Wrapping绕回方式：在故障链路两端的节点内部把0环和1环连接在一起，重新形成一个新的环网。
- Steering抄近方式：更新拓扑结构，重新选路。

RPR环的保护倒换（续）

对于环上正在传送的数据流量，存在两种保护方式：绕回（wrapping）保护方式和抄近(steering)保护方式。

Wrapping保护方式下，当环上某个地方发生故障时，故障附近两个节点处将自动环回，即把内环和外环连在一起，形成一个闭合单环，整个环可利用带宽减少50%。环回后，经由故障节点/段的业务将在环回节点处环回，绕行相反方向，然后在另外一个环回节点处返回到原来方向，并继续传送到目的节点。如图1所示，故障前从F节点到C节点的数据流，走0环，路径为F-A-B-C；当A节点和B节点之间的链路故障后，采用绕回保护方式，在故障链路两端的节点上通过光路环回，数据路径也在此环回，总的路径为F-A-F-E-D-C-B-C。Wrapping操作非常快，几乎没有包损失，但是Wrapping后，由于业务包在环上绕行，带宽有所浪费，特别是故障临近的段，对并发业务影响较大，难免发生业务拥塞的现象。

 Page27 

RPR环的保护倒换（续）

Steering 保护方式下，当环上某个地方发生故障时，指名故障点和类型的Steering保护消息会瞬时发送到环上每个节点，拓扑也会相应更改。有了新的拓扑，源节点只需要直接按新的拓扑发送数据给目的节点即可，由于路径选择是根据新的拓扑做出的，数据可以经由一个方向直接到达目的节点，无需从发生故障的地方环回。如图2所示，故障前从F节点到C节点的数据流，走0环，路径为F-A-B-C；当A节点和B节点之间的链路故障后，采用抄近保护方式，从F节点到C节点的数据流量改抄近道，走另外一个环（1环）到达目的节点，路径为F-E-D-C。Steering保护方式下，原来的双环结构变化为非闭合的两个开口的单环，可用带宽同样减少到50%，而且由于Steering操作稍慢一点，在新拓扑获得以前，已经发出的小部分数据将在故障点被丢弃（开环）。Steering保护方式下，数据没有绕行，不会由于绕行而浪费带宽。

 Page28



两种方式的比较：

- 绕回方式的优点是故障切换的恢复时间非常短(50ms以内)，只可能丢失极少量的报文，不会造成业务中断的情况，问题是占用带宽较多。
- 抄近方式避免了带宽的浪费，但是由于需要重新收敛，恢复时间较长，可能会造成一些业务的中断。

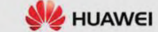
华为公司同时支持绕回和抄近这两种方式，并且结合这两种方式各自的优点，取长补短，配合应用，采取先用绕回方式，后转用抄近方式的保护流程。当RPR链路故障出现时，立即启动绕回方式进行保护，不中断业务；当各节点的各种状态数据(拓扑信息等)重新收敛并稳定下来后，切换到抄近方式，以便节省带宽。这样可以达到最佳的保护和利用效果。




目 录
环网技术概述
RPR环技术概述
RPR环技术特点
RPR环的基本概念
RPR环的数据操作
RPR环的保护倒换
RPR环的公平算法
Copyright © 2012 Huawei Technologies Co., Ltd. All rights reserved.
Page29
HUAWEI

RPR环的公平算法

RPR通过RPR公平算法或RPR-fa算法（RPR Fairness Algorithm）进行拥塞控制。当一个节点发生拥塞时，它通过反方向的环向上行节点发送RPR的使用报文（Usage Packet），该报文同时还起到维持链路状态的作用，上行节点根据报文中的信息调整自己发送数据的速率，以消除拥塞。

 Page30 

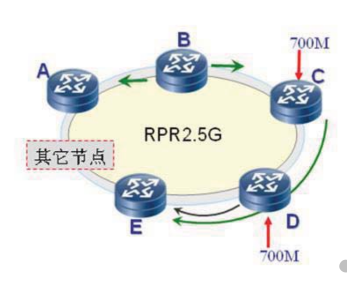
RPR采用共享带宽方式实现各节点对带宽资源的利用。当数据流量较小的情况下，RPR可以满足所有节点流量上载的需求。但是当流量较大的时候，往往会出现链路过载、流量拥塞的情况，流量对链路带宽占用需求不能得到完全满足，在这种情况下，有些节点可以会利用自身位置优势(近水楼台先得月)或时间优势(先入为主)，过多地霸占带宽，影响其他节点对带宽的享用。为了保证在拥塞或超载等情况下各节点能够公平地享用带宽，RPR为此提供专门的公平算法。

RPR的公平算法是一种分布式的公平算法，节点间通过控制报文传递公平算法所需的各项信息，包括允许速率、建议速率、策略指示等。公平算法包括流量统计和策略处理以及处理中的多个阶段，最终实现流量公平分配。

带宽公平和拥塞控制机制属于RPR数据链路层MAC控制子层部分的功能。RPR公平算法只适用于对带宽需要进行争用的业务，即对EIR业务和尽力传送业务起作用。

RPR环的公平算法（续）

如图所示，RPR环中有A、B、C、D、E五个节点，RPR链路带宽为2.5Gbps，流量通过0环传送。首先C、D节点分别发送700Mbps流量至节点，在D-E段共享带宽，D-E段链路消耗的带宽为1.4Gbps，无拥塞。



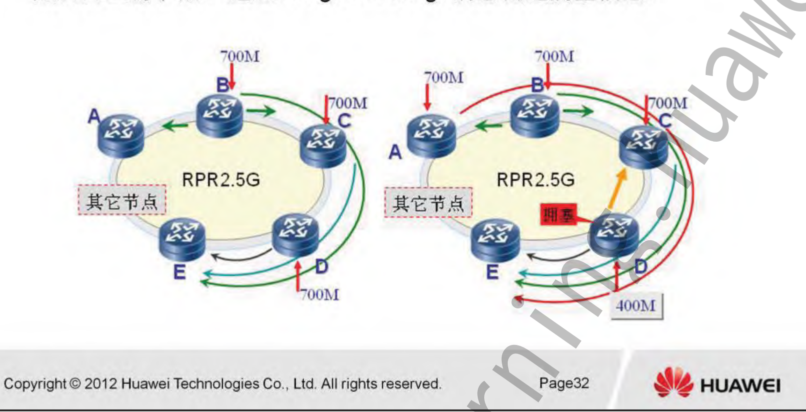
 Page31



1. 节点RouterC、RouterD分别以700Mbit/s的速度向节点RouterE发送数据。RouterC、RouterD发出的数据在C-E 段共享带宽，D-E段链路消耗的带宽为1.4Gbps，无拥塞。

RPR环的公平算法（续）

如图所示，B增加700M后，D-E流量达到2.1G，仍然可以无阻塞转发。但是在A也注入700M流量到E时， D-E上面流量拥塞，4*700M大于2.5G。依据公平算法D节点立刻将本地节点下发流量降为400M，然后向上游节点C 通过Usage Message消息传递拥塞信息。

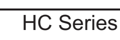


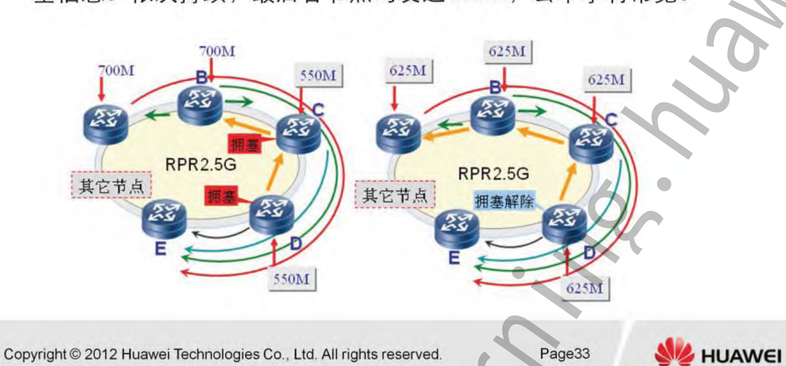
Page32 HUAWEI

2. 节点RouterB也以700Mbit/s的速度向节点RouterE发送数据。D-E流量达到2.1G，仍然可以无阻塞转发。

3.节点RouterA也以700Mbit/s的速度向节点RouterE发送数据。D-E流量达到2.8G，超过RPR 链路带宽极限值2.5Gbit/s，D-E段出现拥塞。

4. 根据公平算法，RouterD进行公平计算，立即将本地上环数据的流量降为400Mbit/s，同时通过1环反向发送控制帧给RouterC，传递拥塞和公平算法信息。





5. 节点RouterC收到控制帧后，立即降低本地上环数据的流量。根据公平算法计算值，RouterD、RouterC两节点的上环数据流量都调整为550Mbit/s。同时节点RouterC继续向上游节点RouterB传送公平算法控制帧。

6. 节点RouterB收到控制帧后也作相应处理。这样依次下来，最后RouterD、 RouterC、 RouterB、RouterA的上环数据流量都调整为625Mbit/s，平均享用带宽。



总 结

本课程主要介绍了以下内容：

传统的环网技术以及RPR环的特点

RPR环的基本概念和数据操作

RPR环的倒换

RPR环的拥塞控制方法--公平算法

Page34



谢谢

www.huawei.com


NSF技术
www.huawei.com
Copyright © 2012 Huawei Technologies Co., Ltd. All rights reserved.
HUAWEI



前 言

系统发生故障时进行主备倒换是提供系统可用性的一种重要方式。主备倒换会导致数据丢失。大部分丢失的数据可通过HSB（Hot Standby）提供的数据平滑过程恢复，对于无法恢复的数据，必须通过GR（Graceful Restart）功能修复。





培训目标

学完本课程后，您应该能：

- 掌握NSF技术的基本概念
- 掌握NSF技术的基本原理

 Page2 HUAWEI


目 录
NSF的基本概念
NSF技术对系统的要求
冗余备份技术
HSB技术
GR技术
Copyright © 2012 Huawei Technologies Co., Ltd. All rights reserved.
Page3
HUAWEI

NSF的基本概念

NSF（None Stop Forwarding，不间断转发）：是一项重要的高可靠性技术，它可以保证路由器控制层面出现故障时，数据转发仍然正常执行，从而保护网络上关键业务不受影响。

通常情况下，路由器故障后，其路由协议层面的邻居会检测到它们之间的邻居关系Down掉，然后过段时间再次Up，这个过程被称之为邻居关系震荡。这种邻居关系的震荡将最终导致路由震荡的出现，使得重启路由器在一段时间内出现路由黑洞或者导致邻居将数据业务从重启路由器处旁路，从而导致网络的可靠性大大降低。不间断转发技术的目标就是为了解决上述路由震荡的问题

 Page4 

对于主控板的冗余备份，当主控板因为硬件或软件失效出现故障时，备用主控板接管失效主控板的工作，重新启动控制平面、管理平面、以及各业务处理板。这个过程一般要持续几分钟，期间数据报文无法处理。

路由器周边的其他路由器通过动态路由协议感知到网络节点故障，重新计算路由。当失效路由器恢复后，路由及各种信令协议重新建立联系。

以上变动将会引发网络的路由振荡。




目 录
NSF的基本概念
NSF技术对系统的要求
冗余备份技术
HSB技术
GR技术
Copyright © 2012 Huawei Technologies Co., Ltd. All rights reserved.
Page5
HUAWEI

NSF技术对系统的要求

硬件要求：系统双主控冗余配置，当主用主控板重启，备用主控板成为新的主板；分布式结构，数据转发和控制分离，有专门的线卡（接口板）用于数据转发。

系统软件要求：主板正常运行的过程中，会把配置信息、接口状态信息备份到备用板；主备倒换的时候，接口板不需要重启，接口保持Up，接口板转发表不撤销。

协议要求：要求各相关网络协议如路由协议OSPF、IS-IS、BGP以及其他协议如LDP、RSVP做扩展以具备GR（Graceful Restart，优雅重启）能力。

 Page6 

在分布式处理模型下，路由器主要由主控板和接口板组成，主控板负责路由器的控制平面和管理平面，例如路由学习、路由计算、建立MPLS LSP等工作；接口板负责具体业务处理，如IP报文转发，MPLS标签交换，QoS保证等工作。

高端路由器的特点是主控板具有冗余备份机制，即双主控板设计：其中一块是主用主控板（AMB），处于工作状态，另一块称作备用主控板（SMB），处于备份状态。主用主控板运行过程中，将所有静态配置信息和一部分动态信息备份到备用主控板，使得备用主控板具有和主用主控板相同的配置信息。当主用主控板因为硬件或软件失效出现故障时，备用主控板接管失效的主用主控板的工作，重新启动控制平面和管理平面工作。

系统具备控制与转发分离的分布式结构，所以主备倒换过程中接口板不会重启，接口板上的转发表被保留，可以继续业务转发。




目 录
NSF的基本概念
NSF技术对系统的要求
冗余备份技术
HSB技术
GR技术
Copyright © 2012 Huawei Technologies Co., Ltd. All rights reserved.
Page7
HUAWEI

冗余备份技术

对系统中关键组件进行冗余备份是提供系统容错能力的主要方法。

冗余备份方式有1+1备份和n+1备份。

1+1备份方式：两个组件必须互为镜像（mirror）。当主用组件发生故障时，备用组件可以立即接管当前任务，从而保证系统业务不中断。

n+1备份方式：系统通过n个相似的组件来提供某项服务，由另外一个组件作为这n个组件的备份。当n个中的某一个组件发生故障时，备份组件接管故障组件的任务，业务仍然可以继续进行。

冗余备份是系统进行主备倒换的前提条件。

 Page8 

目前华为路由器提供的硬件备份功能如下：

- 交换网板之间形成1+1或者n+1备份
- 各主控板（简称为主板）之间形成1+1备份
- 各业务处理板之间形成1+1备份或n+1备份
- 各电源之间形成1+1或者n+1备份
- 各制冷风扇之间形成n+1备份




目 录
NSF的基本概念
NSF技术对系统的要求
冗余备份技术
HSB技术
GR技术
Copyright © 2012 Huawei Technologies Co., Ltd. All rights reserved.
Page9
HUAWEI

HSB技术

HSB相关部件和术语如下：

AMB（Active Main Board）：主用主板。

SMB（Standby Main Board）：备用主板。

HA Channel：主用主板与备用主板之间进行通信的通道。

Switchover：主备板切换，即由管理员或严重故障触发引起的主板切换动作。在此过程中，原主用主板会被复位而变成备用主板。

Smooth：数据平滑过程，即当备用主板切换成主用主板后，新主用主板上不同模块之间的数据可能不一致，此时需要进行数据同步操作使其一致。

Page10



可以进行手工切换主用主板，华为交换设备主备倒换命令：

[Quidway] slave switchover

HSB技术（续）

HSB（Hot Standby）是提供热备份的一个关键技术。HSB的功能是将系统的静态和动态配置信息从AMB备份到SMB。

当系统重启时，主用主板AMB将其静态配置信息备份到备用主板SMB上，在系统正常运行时，AMB上的任何数据变化（包括静态和动态的数据变化）都会备份到SMB。备用主板SMB切换成主用主板AMB后进入平滑运行阶段。切换后，主用主板AMB上的所有数据都已备份，因此，该系统与其它设备间的会话不受影响，其它设备也不会察觉该CX设备的切换。

 Page11 

在系统正常运行时， AMB可将路由信息从数据层面下载到接口板，SMB却不能、也无法从接口板接收任何信息。

主备用主控板之间备份的信息：

- 同步配置更改
- TCP连接
- 接口状态信息
- 各种事件
- 路由/mpls协议状态

主备用主控板之间定期进行心跳检测。

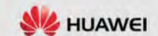

目 录
NSF的基本概念
NSF技术对系统的要求
冗余备份技术
HSB技术
GR技术
Copyright © 2012 Huawei Technologies Co., Ltd. All rights reserved.
Page12
HUAWEI

GR技术概述

IETF针对IP/MPLS转发相关的协议（如OSPF、IS-IS、BGP、LDP和RSVP）进行扩展，实现协议重启时转发不中断的功能，使系统进行主备倒换时控制层协议的震荡在一定程度上得到限制。这一系列标准统称为各个协议的Graceful Restart扩展，简称GR。GR目前已经被广泛的使用在主备倒换和系统升级方面。

系统能够进行GR的前提条件是转发和控制分离，即设备有主控板和接口板，接口板处理转发任务。当系统进行协议重启或者主备倒换时，不复位接口板，接口板继续转发数据，从而实现整个系统不间断地转发报文。

由此可以看出，系统实现不间断转发的必要条件是在GR Time时间内网络拓扑和接口状态不发生变化，否则系统将退出GR过程，转发也将中断。

 Page13 

目前，各种路由协议只运行于路由器主控板上，在备板上并不运行，当主备倒换发生后，备板上开始运行路由协议。

路由协议支持GR能力，需要完成下面两项任务：

- 邻居路由器与重启路由器的邻居关系避免在重启的时候震荡（flap）；
- 重启后，重启路由器尽快完成与邻居路由器的路由信息的同步，然后更新本地路由信息。

目前具备GR能力的路由协议主要有IS-IS、OSPF、BGP。

GR技术基本概念

角色：

GR Restarter：GR重启设备，指路由协议使能了GR能力，能够在主备倒换的时候通知邻居，请求邻居保持与自己的邻接关系。

GR Helper：GR Restarter的邻居，至少能够识别GR信令，在GR Restarter进行主备倒换时保持和GR Restarter的邻接关系不变，协助GR Restarter进行网络拓扑关系的恢复。

会话和定时器：

GR Session：有GR能力的会话，是GR Restarter和GR Helper之间通过GR能力协商建立的会话关系。

GR Time：是GR Helper发现GR Restarter Down后，保持路由信息不删除的时间，可以看作从GR过程开始到结束这一段时间。

 Page14 

GR Restarter与GR Helper的作用是相互的。在GR Helper使能GR能力的情况下，GR Restarter与GR Helper的位置和作用可以互换。

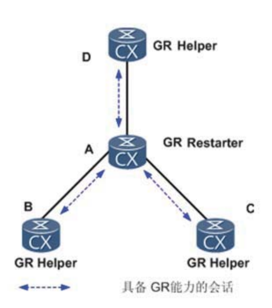
各协议GR实现机制不一致， GR Time 也不一致：

- OSPF缺省情况下，重启的时间为120秒。
- IS-IS缺省情况下，重启的时间为300秒。
- BGP缺省情况下，重启的时间为150秒。
- MPLS LDP缺省情况下，会话重连接定时器的值为300秒。



GR的工作过程

Restarter和GR Helper之间进行GR能力协商，建立会话 ，如下图所示，A作为GR Restarter，B、C和D分别是A的GR Helper邻居，在GR Restarter和GR Helper之间建立起有GR能力的会话



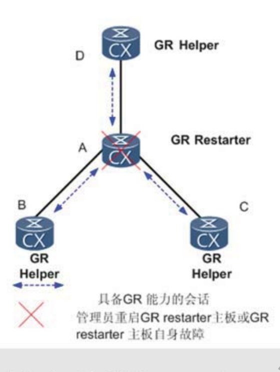
 Page15 

前提条件：

- 路由器配置有双主控，转发和控制分离。
- 路由协议已经达到正常状态。
- 要求各相关网络协议（如路由协议OSPF、IS-IS、BGP，其他协议如LDP、RSVP做扩展），具备优雅重启（GR）能力。
- 配置好GR。

GR的工作过程（续）

GR Helper发现GR Restarter故障，继续保持和GR Restarter的邻接关系，在GR Time超时之前，仍保留与GR Restarter相关的路由



　Page16　

以下为针对链路状态算法，以IS-IS为例：

路由器发生主备倒换后，由于没有保存任何重启前邻居信息，因此一开始发送的Hello报文中不包含邻居列表，此时邻居路由器收到后，执行双向邻居关系检查，发现在重启路由器的Hello报文的邻居列表中没有自己，这样邻居关系将会断掉，同时通过生成新的LSP报文，将拓扑变化的信息泛洪给区域内的其他路由器。这样区域内路由器基于新的链路状态数据库进行路由计算，从而造成路由中断或者路由旁路。

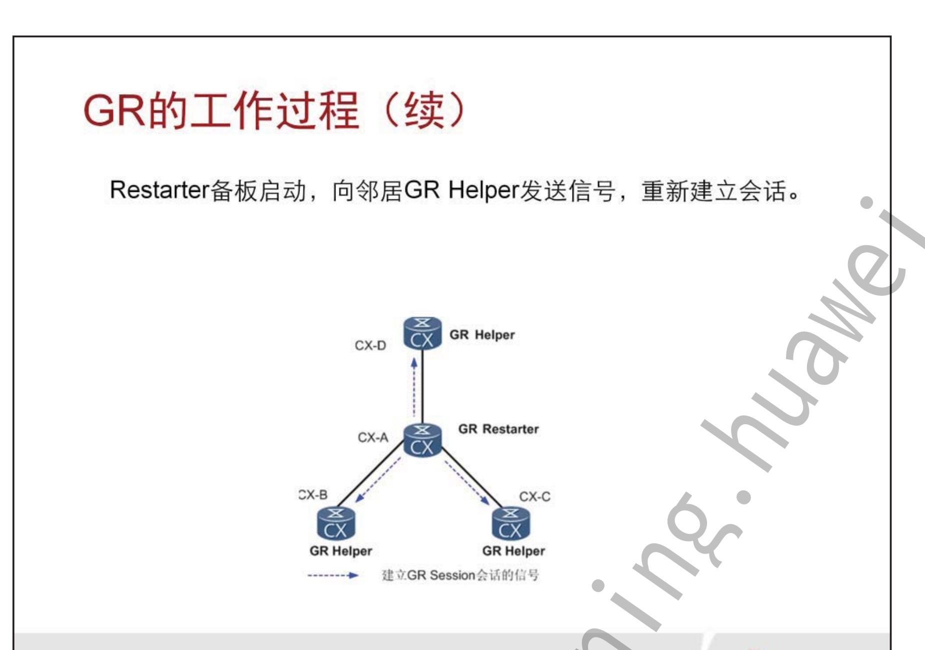
由于没有保存重启前的任何链路状态信息（LSDB），重启路由器在主备倒换后，需要快速和邻居间同步链路状态信息。

为此，IS-IS为了支持GR能力，需要完成下面两项任务：

- 1）GR-Capable路由器重启期间，要避免邻居路由器针对重启路由器的邻居关系的震荡，即GR-Helper对于重启前就处于Up状态的重启路由器邻居关系，在GR期间，仍然保持处于Up状态。为了实现这个功能，IS-IS 在Hello报文中新增加了一个新的TLV 211。TLV211中主要包含RR bit, RA bit。 当RR Bit被设置，表示该路由器刚刚发生了优雅重启（GR），当RA bit = 1 表示是对重启路由

器的应答。另外TLV211中还包含一个Remaining time字段，是邻居路由器在应答Hello报文中携带的，表示邻居能容忍的重启路由器重启所要消耗的最大时间。

- 2）在邻居关系保持的基础上，重启发生后，新的主板上的IS-IS需要尽快同各个具备GR-Aware能力的邻居同步链路状态数据库。具备GR-Aware能力的邻居路由器在收到重启路由器的GR请求后，需要向重启路由器同步它所具备的链路状态数据库，具体做法是：将所有的LSP都打上向重启路由器发送标志；向重启路由器发送完全数据库列表CSNP报文。



IS-IS GR 过程描述：

- 1、CX-A路由器重启前，使能IS-IS的GR能力，这时候IS-IS发出的Hello报文中携带TLV211，但是RR bit和RA bit都被置为0。
- 2、CX-A路由器重启，发生主备倒换。接口板不重启，接口板上转发表的路由被打上老化标志，但是仍然继续正常指导转发。
- 3、CX-A原来的备板成为新的主板。IS-IS协议重新启动，从系统得知这是一次主备倒换后的重启，并且如果IS-IS提前配置了GR能力的话，则启动T1，T2，T3定时器。
- 4、连接CX-B的接口的T1定时器超时触发，CX-A向CX-B发送一个包含TLV 211的Hello报文，其中RR bit = 1, RA bit = 0，通知邻居 该路由器刚刚发生了重启。
- 5、具备GR-Aware能力的CX-B收到了R1表达重启的Hello消息后，发送同样包含TLV211的Hello报文做应答，其中RR bit = 0, RA bit = 1，表示响应R1的GR重启要求。同时将TLV 211中的Remaining Time字段填充为本系统Hold Time时间。
- 6、如果CX-A和CX-B之间是P2P连接；或者如果是广播网且CX-B的 System id是该广播网上最大的（除了CX-A），则向CX-A发送CSNP消息。CSNP报文是CX-B用来向CX-A同步链路状态数据库

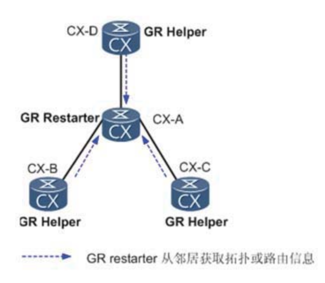
（LSDB）用的，里面包含了CX-B完整的链路状态摘要信息列表。同时，CX-B将所有的LSP设上SRM标志，等待定时器调度，向CX-A发送。

- 7、CX-A收到CX-B含RA Bit = 1的Hello应答，并且收到邻居发送来的CSNP消息后，则T1定时器被删除。否则，T1定时器触发继续向CX-A发送RR bit = 0, RA bit = 1的Hello报文，发送n次后，T1被删除。
- 8、如果CX-A同步到了所有GR-Aware邻居的LSDB（通过检查记录的邻居发来的CSNP链路状态摘要信息列表最后是否被清空判断得知），那么T2，T3定时器都被删除，执行10。注意，在此之前，CX-A可以生成自己的LSP，但是不能发送出去，收到邻居发来的自己产生的LSP也不能清除（Purge），也不能进行路由计算，原因在于基于没有完全同步的 LSDB的计算可能导致路由错误或者丢失。如果T2定时器超时，执行10。
- 9、T3定时器被触发，表明R1没有在GR-Aware邻居允许的时间内完成LSDB同步，那么将在自己生成的 LSP中设置Overload bit后，发送给邻居。
- 10、调度路由计算，路由计算结束后，更新接口板转发表；清除掉重启过程中可能收到的非法的LSP，将自己生成的LSP 泛洪出去；IS-IS协议恢复正常流程。

GR的工作过程（续）

GR Restarter从邻居获取拓扑和路由信息后，重新计算路由表，并老化旧路由。

至此，GR Restarter完成主备切换过程，并且在主备倒换期间转发不中断。







支持GR的协议

目前VRP支持GR能力的协议有：

MPLS LDP

OSPF（IPv4）

IS-IS（IPv4/IPv6）

BGP（IPv4/IPv6/VPNv4）以及带标签路由BGP

RSVP

 Page21 

路由协议OSPF、IS-IS、BGP，其他协议如LDP、RSVP都做了扩展，以支持优雅重启（GR）能力。



总 结

本课程主要介绍以下内容：

NSF的基本概念和对系统的要求

冗余备份技术、HSB技术、GR技术以及支持GR的协议

 Page22



谢谢
www.huawei.com


快速检测技术
www.huawei.com
Copyright © 2012 Huawei Technologies Co., Ltd. All rights reserved.
HUAWEI



前 言

为了减小设备故障对业务的影响，提高网络的可用性，网络设备需要能够尽快检测到与相邻设备间的通信故障，以便及时采取措施，保证业务继续进行。

 Page1



网络设备一个越来越重要的趋势是，要求对相邻系统之间通信故障进行快速检测，这样在出现故障时可以更快的建立起替代通道或倒换到其他链路。



培训目标

学完本课程后，您应该能：

- 快速检测技术的基本概念
- 快速检测技术的基本原理和应用

 Page2 HUAWEI


目 录
故障检测的主要方法
故障检测的分类
BFD技术
Copyright © 2012 Huawei Technologies Co., Ltd. All rights reserved.
Page3
HUAWEI

故障检测的主要方法

现有的故障检测方法主要包括：

硬件检测：例如通过SDH告警检测链路故障。硬件检测的优点是可以很快发现故障，但并不是所有介质都能提供硬件检测。

慢Hello机制：通常是指路由协议的Hello机制。这种机制检测到故障所需时间为秒级。对于高速数据传输，例如吉比特速率级，超过1秒的检测时间将导致大量数据丢失；对于时延敏感的业务，例如语音业务，超过1秒的延迟也是不能接受的。并且，这种机制依赖于路由协议。

其他检测机制：不同的协议或设备制造商有时会提供专用的检测机制，但在系统间互联互通时，这样的专用检测机制通常难以部署。

　Page4



很多硬件或者软件无法提供硬件检测功能，比如以太网。还有一些无法实现路径检测，比如转发引擎或者接口等，无法实现端到端的检测。

慢Hello机制：对于不允许路由协议的节点没有办法检测链路的状态。

用慢Hello机制，尤其在路由协议中，在没有硬件帮助下，检测时间会很长（例如：OSPF需要2秒的检测时间，ISIS需要1秒的检测时间）。




目 录
故障检测的主要方法
故障检测的分类
BFD技术
Copyright © 2012 Huawei Technologies Co., Ltd. All rights reserved.
Page5
HUAWEI

故障检测的分类

故障检测技术按使用限制分专用检测技术和通用检测技术：

专用的故障检测技术有：

- APS（传输层）
- RPR OAM、Eth-OAM（链路层）
- MPLS OAM（MPLS）

通用故障检测技术包括：BFD，可检测各个层面的故障。

 Page6



自动保护倒换APS：

- 基于底层光网络的方案，如利用SDH的APS功能，利用WDM的波长保护功能 。利用自动保护倒换机制（APS）网络，可以在检测到一条路径故障之后自动将业务流倒换到另一条路径上。APS具体实现可以采用1+1，1:1，或1:N保护方式。

MPLS OAM技术为MPLS网络提供了一套缺陷检测的工具及缺陷纠正机制，通过MPLS OAM及保护倒换构件可以完成CR-LSP转发平面的检测功能，并在缺陷发生后的50ms内完成保护倒换，从而将缺陷所产生的影响减小到最低。



故障检测的分类（续）

以TCP/IP网络参考模型作为层次性划分，每个层面都有故障检测机制:

传输层/物理层：APS

链路层：RPR OAM、MPLS OAM、Eth-OAM、STP/RSTP/MSTP/RRPP

网络层：各协议的HELLO机制、BFD、VRRP、GR

应用层：各种应用层协议本身的心跳、重传机制

 Page7 

传输层TCP的故障检测机制：重传机制、滑动窗口机制。

故障检测的分类（续）

以网络故障检测的模式划分，有以下3种模式：

异步模式：周期性发送探测报文

查询模式：发一系列报文请求确认

回声模式：将对端发送过来的报文不作任何改动反射回去

 Page8



异步模式：系统之间相互周期性地发送BFD控制包，如果某个系统在检测时间内没有收到对端发来的BFD控制报文，就宣布会话为Down。

查询模式：，假定每个系统都有一个独立的方法用来确认它连接到其他系统。这样一旦一个会话建立起来以后，系统停止发送控制报文，除非某个系统需要显式地验证连接性，在需要显式验证连接性的情况下，系统发送一个短系列的控制包，如果在检测时间内没有收到返回的报文就宣布会话为Down，如果收到对端的回应报文，协议再次保持沉默。

回声功能：本地发送一系列回声报文，远端系统通过它的转发通道将它们环回回来。如果本地系统连续几个回声报文都没有接收到，会话就被宣布为Down。回声功能可以和上述两种检测模式一起使用，可以使用回声功能来代替控制报文的检测的任务，这样可以降低控制报文的发送周期（异步模式下）或者完全取消控制报文（查询模式下）。

异步模式和查询模式的本质区别在于检测的位置不同，异步模式下本端按一定的发送周期发送控制报文，需要在远端检测本端系统发送的控制报文；而在查询模式下检测本端发送的控制报文是在本端系统进行的。




目 录
故障检测的主要方法
故障检测的分类
BFD技术
Copyright © 2012 Huawei Technologies Co., Ltd. All rights reserved.
Page9
HUAWEI

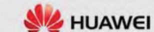
BFD技术概述

双向转发检测BFD（Bidirectional Forwarding Detection）是一套全网统一的检测机制，用于快速检测、监控网络中链路或者IP路由的转发连通状况。为改善网络性能，相邻系统之间应能快速检测到通信故障，更快地建立起备用通道恢复通信。

BFD提供如下功能：

- 对相邻转发引擎之间的通道故障提供轻负荷、短持续时间的检测；
- 用单一的机制对任何介质、任何协议层进行实时检测，并支持不同的检测时间与开销。

 Page10



BFD是一个简单的“Hello”协议，在很多方面，它与那些著名的路由协议的邻居检测部分相似。一对系统在它们之间的所建立会话的通道上周期性的发送检测报文，如果某个系统在足够长的时间内没有收到对端的检测报文，则认为在到相邻系统的双向通道的某个部分发生了故障。在某些条件下，为了减少负荷，系统之间的发送和接收速率需要协商。

BFD(双向转发检测)是一套用来实现快速检测的国际标准协议，提供一种轻负荷、持续时间短的检测。与以往的其他”HELLO”检测机制相比，具有许多独到的优势。

BFD能够在系统之间的任何类型通道上进行故障检测，这些通道包括直接的物理链路，虚电路，隧道，MPLS LSP，多跳路由通道，以及非直接的通道。

BFD通过在双向链路两端同时发送检测报文，检测两个方向上的链路状态，实现毫秒级别的链路故障检测。

双向链路的一种特殊情况是单向链路，例如 LSP，这时只需在一个方向发送 BFD控制报文，对端通过其他路径报告链路状况。



BFD for IP技术

在IP链路上建立BFD会话，利用BFD检测机制快速检测故障。

BFD for IP支持单跳检测和多跳检测:

BFD单跳检测是指对两个直连系统进行IP连通性检测，这里所说的“单跳”是IP的一跳。在进行BFD单跳检测的两个系统中，对于一种给定的数据协议，在指定接口上只存在一个BFD会话。

BFD多跳检测是指BFD可以检测两个系统间的任意路径，这些路径可能跨越很多跳，也可能在某些部分发生重叠。

 Page11



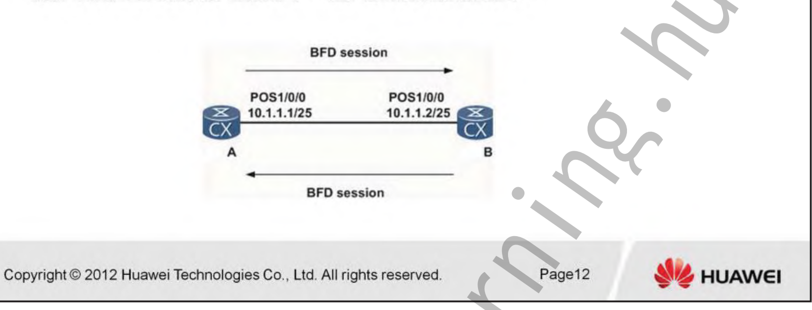
单跳检测采用一种简单的方法来防止遭到欺骗（Spoofing）攻击。发送 BFD控制报文时，携带的 TTL 或跳数值必须是 255。如果收到 TTL 不是 255 的 BFD控制报文，必须将这些报文丢弃。

在多跳检测中，封装 BFD控制报文的 UDP 报文的目的端口号是 3784，源端口号取值范围是 49152～65535。

最新的BFD草案中，多跳检测的UDP目的端口号是4784。

BFD for IP技术（续）

如下图所示， BFD单跳检测是指对两个直连系统进行IP连通性检测，这里所说的“单跳”是IP的一跳。在进行BFD单跳检测的两个系统中，对于一种给定的数据协议，在指定接口上只存在一个BFD会话。因此，BFD会话是与接口绑定的，接口类型包括物理接口、虚电路以及隧道。



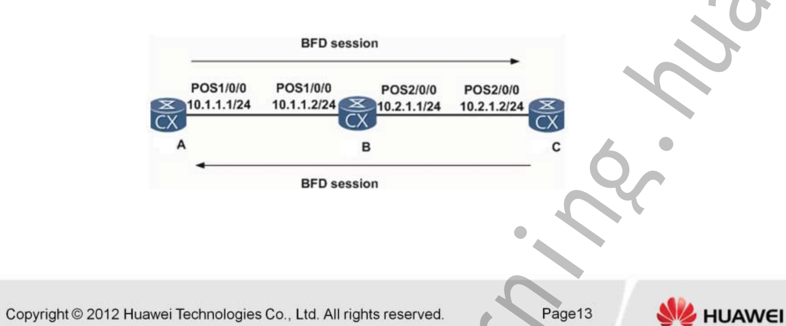
　Page12　HUAWEI

BFD的检测机制是两个系统建立 BFD会话，并沿它们之间的路径周期性发送 BFD控制报文，如果一方在既定的时间内没有收到BFD控制报文，则认为路径上发生了故障。

双向链路的一种特殊情况是单向链路，例如 LSP，这时只需在一个方向发送 BFD控制报文，对端通过其他路径报告链路状况。

BFD for IP技术（续）

如下图所示， BFD可以检测两个系统间的任意路径，这些路径可能跨越很多跳，也可能在某些部分发生重叠。 多跳BFD会话绑定对端IP但不绑定出接口。



 Page13 HUAWEI

BFD封装在UDP中。在路由器B上要对BFD报文进行解封装、封装、路由等操作，与普通数据包的转发一样。

BFD for USR技术

BFD for USR（Unicast Static Route）用于支持IPv4单播静态路由，支持IPv4单播静态路由绑定后快速感知链路状态。

与动态路由协议不同，单播静态路由自身没有检测机制，当网络发生故障的时候，需要管理员介入。BFD for USR特性可为公网IPv4单播静态路由绑定BFD会话，利用BFD会话来检测单播静态路由所在链路的状态。

BFD for USR可为每条IPv4单播静态路由绑定一个BFD会话，当这条USR上绑定的BFD会话检测到链路故障（由Up转为Down）后，BFD会将故障上报路由管理系统，由路由管理模块将这条路由设置为“非激活”状态（此条路由不可用，从IP路由表中删除）。

当这条USR上绑定的BFD会话成功建立或者从故障状态恢复后（由Down转为Up），BFD会上报路由管理模块重新激活。

 Page14 

使用静态路由Track静态BFD会话的功能可以实现故障检测和路由快速收敛。将两个节点之间的静态路由与BFD会话绑定，当BFD检测到故障时，可以促使静态路由快速失效，以减少对上层业务的影响。

BFD for IGP技术

通常情况下，IGP设定发送Hello报文的时间间隔为十几秒钟，宣告邻居Down的时间即相邻设备失效的时间一般配置为Hello报文间隔的3-4倍。通过调整Hello报文间隔，设备能感知到邻居故障的时间最小也是秒级。在高速的网络环境中，这将导致报文大量丢失。

BFD for IGP是指BFD会话由IGP协议动态创建，不再依靠手工配置，当BFD检测到故障时，通过路由管理通知IGP协议，由协议进行相应邻居Down处理，快速更新路由信息和进行增量路由计算，从而实现路由的快速收敛。

　Page15



BFD使用本地标识符（Local Discriminator）和远端标识符（Remote Discriminator）区分同一对系统之间的多个BFD会话。IS-IS协议支持动态和静态方式建立BFD会话，OSPF协议支持动态建立BFD会话。

路由协议动态触发建立BFD会话的实现方式是：

- 动态分配本地标识符
- 自学习远端标识符

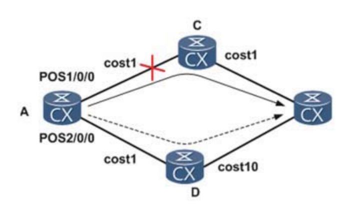
当路由协议邻居建立成功时，路由协议通过路由管理模块通知BFD建立会话，对路由协议的邻居关系进行快速检测。BFD会话的检测参数由路由协议设置。

当BFD会话检测到故障时，状态变为Down，BFD通过路由管理模块触发路由收敛。

当邻居状态不可达时，路由协议通过路由管理模块通知BFD删除相应会话。

BFD for IGP技术（续）

如图所示，当A和C之间链路出现故障，BFD首先感知到并通知A。A处理邻居Down事件，重新进行路由计算，新的路由出接口为POS2/0/0，经过D到达B。



 Page16



BFD for静态路由：当路由器之间有传输设备、交换机、防火墙等设备时，由于静态路由自身没有检测机制，无法感知链路变化，不能动态收敛。使用静态路由Track静态BFD会话的功能可以实现故障检测和路由快速收敛。将两个节点之间的静态路由与BFD会话绑定，当BFD检测到故障时，可以促使静态路由快速失效，以减少对上层业务的影响。

BFD能够为IS-IS/OSPF邻居之间的链路提供快速检测功能，当邻居之间的链路出现故障时，加快IS-IS/OSPF协议的收敛速度。

BFD for IS-IS可以通过静态或动态方式建立BFD会话。

BFD for OSPF可以通过动态方式建立BFD会话。



BFD for BGP技术

BGP协议通过周期性的向对等体发送Keepalive报文来实现邻居检测机制。但这种机制检测到故障所需时间比较长，超过1秒钟。当数据达到吉比特速率级别时，将会导致大量的数据丢失，从而无法满足电信级网络高可靠性的需求。

因此，BGP协议通过引入BFD for BGP特性，利用BFD的快速检测机制，迅速发现BGP对等体间链路的故障，并报告给BGP协议，从而实现BGP路由的快速收敛。

 Page17



BFD能够为BGP邻居之间的链路提供快速检测功能，当邻居之间的链路出现故障时，加快BGP协议的收敛速度。

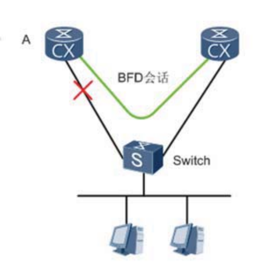
BGP协议支持动态建立BFD会话。

路由协议动态触发建立BFD会话的实现方式是:

- 动态分配本地标识符
- 自学习远端标识符

BFD for VRRP技术

VRRP设定发送心跳报文的时间间隔为1秒钟，宣告邻居Down的时间是心跳报文间隔的3倍。设备能感知到邻居故障的时间最小也是秒级。VRRP通过监视BFD会话状态实现主备快速切换，切换时间控制在50毫秒以内。



Page18



VRRP（Virtual Router Redundancy Protocol）虚拟路由冗余协议，来解决局域网主机访问外部网络的可靠性问题。VRRP是一种容错协议，它通过把几台路由设备联合组成一台虚拟的路由设备，并通过一定的机制来保证当主机的下一跳路由器出现故障时，可以及时将业务切换到其它路由器，从而保持通讯的连续性和可靠性。

在基于VRRP的可靠性组网中，BFD为主备路由器之间的链路提供快速检测机制，当检测到链路故障时，会通告VRRP模块，以实现主备路由器之间的快速切换功能。

BFD能够对接入以太网段或独立的网段进行故障检测。通过与路由、传输及隧道系统中的倒换机制配合，在发生故障时，快速触发倒换，保证网络的可靠运转。



BFD for LSP技术

BFD可以用来检测MPLS LSP转发路径上数据平面的故障。

检测MPLS LSP的连通性时，BFD会话协商有两种方式：

- 静态配置BFD：通过手工配置BFD的本地标识符和远端标识符，由BFD本身的协商机制建立会话。
- 动态创建BFD会话：通过在LSP Ping报文中携带BFD Discriminator TLV进行会话协商。

BFD使用异步模式检测LSP的连通性，即Ingress和Egress之间相互周期性地发送BFD报文。如果任何一端在检测时间内没有收到对端发来的BFD报文，就认为LSP状态为Down，并向LSPM上报LSP Down消息。

 Page19



BFD 可以用来检测MPLS LSP 转发路径上数据平面的故障，同时 BFD 的报文格式是固定的，非常适合在硬件上实现和穿越防火墙。因此用BFD 检测MPLS LSP 数据平面的故障。

具有以下优势：

- 反向只要求IP 路由可达
- 快速检测
- 支持大规模LSP 的故障检测

BFD for LSP技术（续）

BFD支持检测的LSP类型有:

静态BFD for静态LSP

静态BFD for LDP LSP

动态BFD for LDP LSP

HUAWEI

BFD for PST技术

当BFD检测到故障时，修改端口状态表PST（Port State Table）中的接口状态，从而触发快速重路由。BFD会话修改端口状态表功能只能用于绑定接口的BFD单跳会话。

BFD for PST在很多类型的FRR（快速重路由）中使用广泛。在绑定接口的BFD会话中使用BFD for PST，会将该BFD会话与这个接口的PST表联动。在BFD会话检测到链路Down后，将该接口的PST表对应比特位置Down，从而立即进行FRR切换。

　Page21



对于采用硬件转发的设备，直接触发该端口上的业务切换；对于采用软件转发的设备，应用程序通过查询 PST 了解端口状态。

如果允许BFD修改端口状态表PST（Port State Table），当BFD检测到接口状态变为Down时，将更改PST中相应表项。这样，其他上层应用协议就能够通过PST了解接口是否发生故障。

目前，对于NE80E/40E，基于BFD检测的LDP FRR和IP FRR需要通过PST来感知BFD检测结果。

BFD for PIS技术

BFD for PIS（Process interface status）提供一种简单的机制，使得BFD检测行为可以关联接口状态，提高了接口感应链路故障的灵敏度，减少了非直连链路故障导致的问题。

BFD的PIS机制，对检测到链路故障的BFD会话，会立即上报Down消息到相应接口，使得接口进入一种特殊的Down状态：BFD Down状态，该状态等效于链路协议Down状态，在该状态下只有BFD的报文可以正常处理，从而使接口也可以快速感知链路故障。

对于每个要配置接口联动的BFD会话，配置为组播检测并指定接口方式，从而避开对接口IP属性的依赖性。

 Page22



当链路中间存在传输设备时，虽然三层仍是直连，但由于实际物理线路分段，一旦链路故障，两端设备需要比较长的时间才能检测到，导致直连路由失效慢，网络中断时间长。此时，使用BFD for PIS可以解决问题。

组播BFD技术

组播BFD用于检测无IP地址等三层属性的接口之间的链路连通性，达到链路故障快速检测。

通过IP组播发送检测报文，在所需检测链路之间的设备上配置组播检测。本端发送组播报文，如果链路连通，则对端接口也可以收到这个组播报文，上送对端BFD应用，感知链路正常。对于二层Trunk链路，由于发送的是组播报文，IP层转发不需要三层属性，直接下发链路层发送，快速检测链路的连通性。这里的IP是BFD模块配置的公认的组播地址Default-IP，任何收到此IP的接口都将此报文上送BFD应用，完成IP转发。

 Page23



PIM邻居之间动态创建BFD会话检测邻居之间的链路状态，一旦有故障，BFD会把结果直接通告给PIM。



总 结

本课程主要介绍了一下内容:

故障检测的主要方法和分类

BFD技术的基本原理和应用

 Page24



谢谢

www.huawei.com


FRR技术
www.huawei.com
Copyright © 2012 Huawei Technologies Co., Ltd. All rights reserved.
HUAWEI



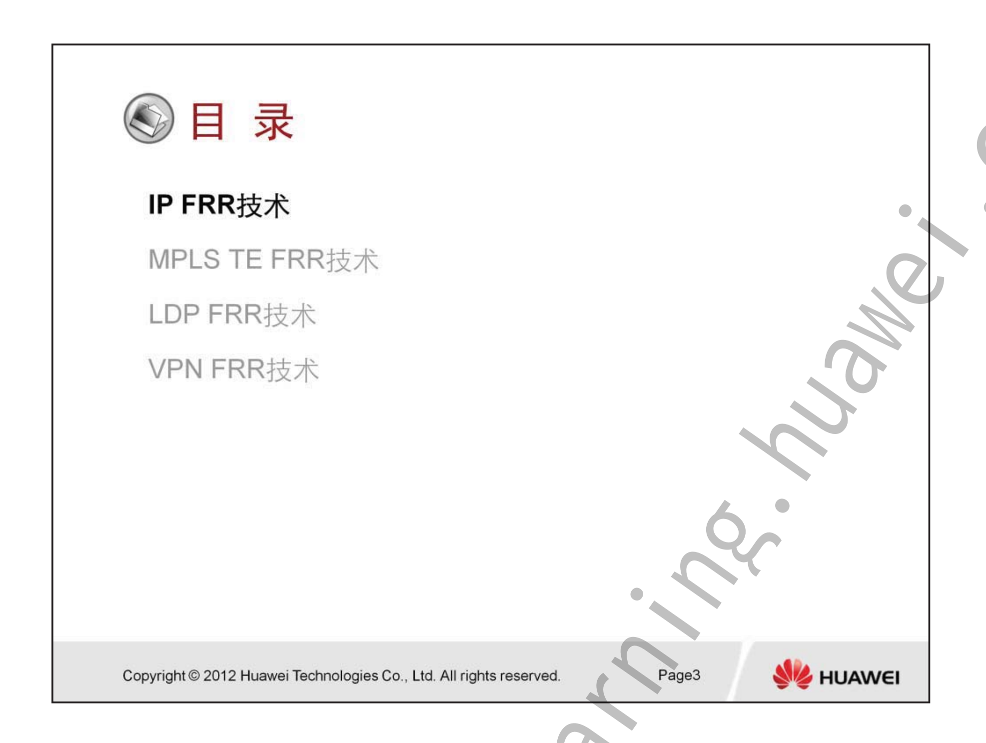
前 言

FRR技术提供了一套快速倒换的机制，通过FRR快速保护倒换结合快速检测技术可以完成IP和MPLS网络的快速故障检测和倒换功能，并在缺陷发生后的50ms内完成保护倒换，从而将缺陷所产生的影响减小到最低。





FRR（Fast ReRoute）是一种可靠性技术。


目 录
IP FRR技术
MPLS TE FRR技术
LDP FRR技术
VPN FRR技术
Copyright © 2012 Huawei Technologies Co., Ltd. All rights reserved.
Page3
HUAWEI

IP FRR技术概述

IP快速重路由针对被保护接口上的IP流量实施快速倒换，速度可达50ms以内。

IP FRR原理是采用一个接口作为另外一个接口的备份，主路径和备份路径都下发到FIB转发表项。主路径没有故障时，流量从主路径发送出去。当主用接口失效，或主用接口连接的邻居失效后，本路由器通过硬件技术或其他快速故障检测协议（如BFD）快速感知，如可通过BFD联动IP FRR，在路由收敛之前将通过这个接口转发的流量快速倒换到备份接口上。

 Page4 

BFD---Bidirectional Forwarding Detection 。

BFD是一个简单的 “Hello”协议，和路由协议的邻居检测部分相似。

- 一对系统在它们之间所建立会话的通道上周期性的发送检测报文，如果某个系统在足够长的时间内未收到对端的检测报文，则认为在这条到相邻系统的双向通道的某个部分发生了故障。

IP FRR是本地保护技术。

IP FRR技术的分类与实现

IP FRR针对IP网络路由而设计，分为公网IP FRR和私网IP FRR：

公网IP FRR：用于保护公网路由器。

私网IP FRR：用于保护CE路由器。

IP FRR的主要实现手段如下：

在主链路可用时，通过Route-Policy设置IP FRR策略，把备份路由的转发信息同时提供给转发引擎。

当转发引擎感知到主链路不可用时，能够在控制平面路由收敛前直接使用备份路径转发信息。

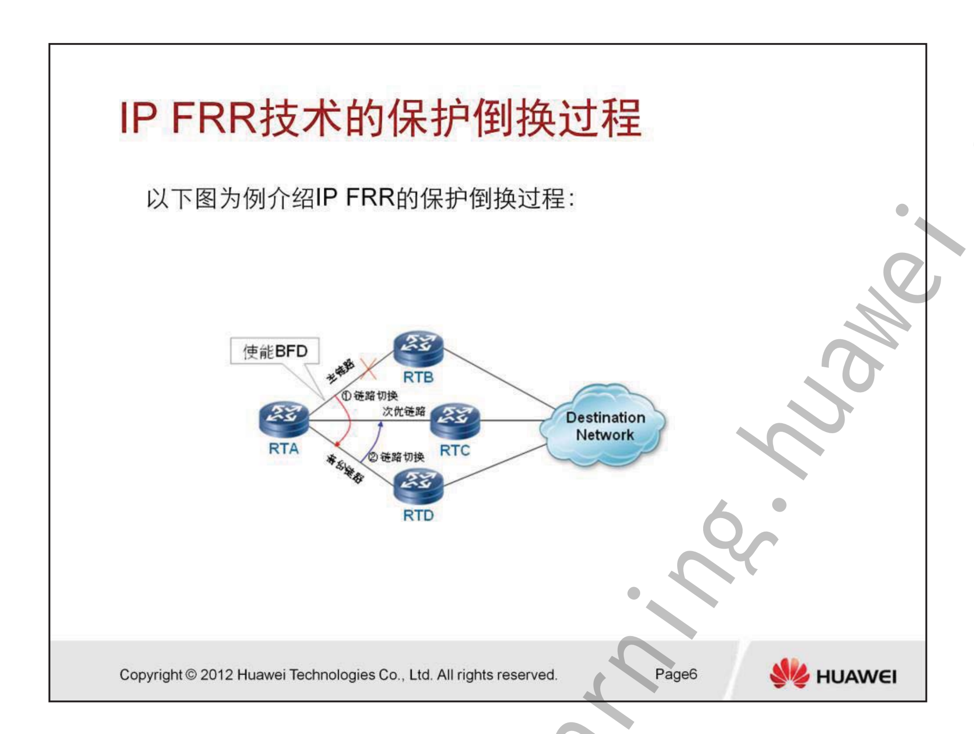
　Page5



在转发模块中建立一张端口状态表，保存设备中每个端口的工作状态。当检测到端口不能正常工作时（如物理链路失效或人工操作将端口关闭），立即更新端口状态表。

同时，在报文转发过程中，如果查转发表得到的表项含有负载分担项（既有多个下一跳），在按照某种规则选定一个下一跳后，在端口状态表中检查这个下一跳出端口的状态，如果状态为失效，则使用另一个下一跳进行尝试，直至遍历全部负载分担项为止。

在检测到最后一个负载分担项时可以不再检查出端口的状态，而直接使用这个下一跳发送报文。



BFD用于检查故障。

故障发生后切换到备份链路。

路由收敛后再切换到次优路由。

也支持负载分担转发。

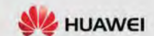
IP FRR技术的保护倒换过程

以上图为例介绍IP FRR的保护倒换过程：

IP FRR技术涉及到主链路、次优链路、备份链路三种链路。主链路指路由最优链路，在网络稳定、路由收敛的情况下，业务量从该链路转发。次优链路指路由cost值比主链路大的链路，主链路失效时，路由会收敛到该链路。备份链路指备份下一跳指定的链路。这三种链路的代价是不等值的。

正常情况，以主链路为出接口的最优路由被选中。当主链路故障后，路由重新收敛，路由表会选中以次优链路为出接口的次优路由。在路由收敛前，通过备份下一跳将流量切换至备份下一跳指定的链路，在路由收敛后，按照路由选择新的链路转发，接口备份使命结束。可见，备份下一跳的作用是填补了路由收敛的时间间隙，通过将流量快速切换到备份下一跳的备份链路，保证业务不中断。

 Page7



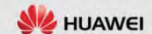

目 录
IP FRR技术
MPLS TE FRR技术
LDP FRR技术
VPN FRR技术
Copyright © 2012 Huawei Technologies Co., Ltd. All rights reserved.
Page8
HUAWEI

MPLS TE FRR技术概述

MPLS TE FRR（MPLS TE Fast Re-Route，MPLS TE快速重路由）是MPLS TE中一套用于链路保护和节点保护的机制。当LSP链路或者节点故障时，在发现故障的节点进行保护，这样可以允许流量继续从保护链路或者节点的隧道中通过，以使得数据传输不至于发生中断，同时头节点就可以在数据传输不受影响的同时继续发起主路径的重建。

MPLS TE FRR的基本原理是用一条预先建立的LSP来保护一条或多条LSP。预先建立的LSP称为快速重路由LSP，被保护的LSP称为主LSP。MPLS TE快速重路由的最终目的就是利用Bypass隧道绕过故障的链路或者节点，从而达到保护主路径的功能。

 Page9



保护的范围为：两个PE设备之间建立端到端的TE隧道 。

是一种局部的保护技术。

MPLS TE是一种将流量工程技术与 MPLS这种叠加模型相结合的技术。通过 MPLS TE，可以建立指定路径的 LSP隧道，进行资源预留，并且可以进行定时优化，在资源紧张的情况下，可以根据优先级和抢占参数的情况，抢占低优先级的 LSP 隧道的带宽资源等等；同时，还可以通过备份路径和快速重路由技术，在链路或节点失败的情况下，提供保护。

Bypass LSP一般处于空闲状态，不承担数据业务。

MPLS TE FRR基本概念

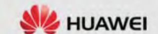
主CR-LSP：被保护的CR-LSP。

Bypass CR-LSP：保护主CR-LSP的CR-LSP。Bypass CR-LSP的入节点是PLR，出节点是MP。Bypass CR-LSP一般处于空闲状态，不承载业务。如果需要使用Bypass CR-LSP保护主CR-LSP的同时承载业务数据的转发，需要为Bypass CR-LSP分配足够的带宽。

PLR（Point of Local Repair）：本地修复节点。Bypass CR-LSP的入节点，必须在主CR-LSP的路径上，可以是主CR-LSP的入节点，但不能是主CR-LSP的出节点。

MP（Merge Point）：汇聚点。Bypass CR-LSP的出节点，必须在主CR-LSP的路径上，并且不能是主CR-LSP的入节点。

 Page10

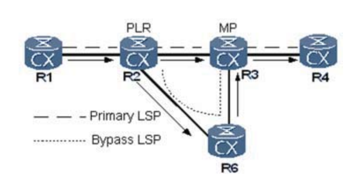


出节点：标签交换路径的末节点。

入节点：标签交换路径的始节点。

MPLS TE FRR基本概念（续）

链路保护：如下图，PLR和MP之间有直连链路（LSR2→LSR3）连接，主CR-LSP经过这条链路。当这条链路失效时，流量可以切换到Bypass CR-LSP（LSR2→LSR6→LSR3）上。

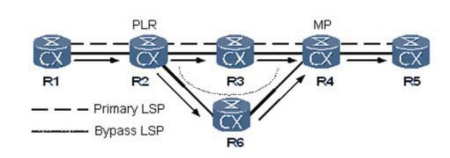


 Page11 HUAWEI



MPLS TE FRR基本概念（续）

节点保护：如下图，PLR和MP之间存在一台LSR（LSR2→LSR3→LSR4），主CR-LSP经过该节点（LSRC）。当该节点失效时，流量可以切换到Bypass CR-LSP（LSR2→LSR6→LSR4）上。



 Page12 HUAWEI





MPLS TE FRR技术的保护倒换过程

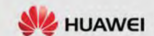
故障检测：

链路保护直接使用链路层协议实现故障检测和通告，链路层发现故障的速度与链路类型直接相关。节点保护则使用链路协议检测链路故障，在链路没有故障的情况下，可以通过配置BFD机制检测被保护节点的故障。

对于节点保护，只保护被保护节点及其与PLR之间的链路。对于被保护节点和MP之间的链路故障，PLR无法感知。

无论是检测到链路故障还是节点故障，最终都会导致PLR上的出接口被置为老化状态。出接口老化就会触发FRR的流量切换。

 Page13



主LSP的建立过程与普通LSP相同。

Bypass LSP的建立可以有两种方式，一种是手工方式，一种是自动方式；

- 手工 Bypass LSP是当一个没有快速重路由属性的隧道被指定保护一个物理接口以后，它所对应的 LSP就成为 Bypass LSP。手工 Bypass LSP的建立是通过在 PLR手工配置触发的。它的配置与普通 LSP基本没有分别，只是不能配置快速重路由属性。也就是说，Bypass LSP不能同时是主 LSP，LSP不能被嵌套保护。
- 自动 Bypass LSP是对手工方式的配置简化，当主 LSP需要被FRR保护时，PLR可以选择或自动建立一条 Bypass LSP，用来保护这个主 LSP，这种方式就叫自动 Bypass。自动 Bypass可以保护多个主 LSP，只要它可以满足这些主 LSP的要求。

Bypass LSP一般处于空闲状态，不承担数据业务。

MPLS TE FRR技术的保护倒换过程（续）

切换：

切换是指主CR-LSP故障后，业务流量和RSVP消息从主CR-LSP切换到Bypass CR-LSP上，并向上游通告切换已经发生。在切换的时候，主CR-LSP的NHLFE表项会被置上切换标志。

 Page14



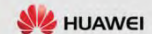
MPLS TE FRR技术的保护倒换过程（续）

回切：

切换后，PLR会向上游发送携带切换标记的PathError消息。Ingress收到该消息，试图重建主CR-LSP，但不拆除主CR-LSP，而是借用Bypass CR-LSP作为主CR-LSP继续转发数据，直到主CR-LSP重建成功。尝试重建的CR-LSP称为Modified CR-LSP。

当主CR-LSP重建成功后，业务流量和RSVP消息需要从Bypass CR-LSP回切到主CR-LSP上。此过程，TE FRR（包括Auto FRR）采用Make-before-break机制，即只有Modified CR-LSP建立成功后，原来的Primary CR-LSP才能被删除。

 Page15



MPLS TE FRR保护的优先顺序

FRR可分为:

带宽与非带宽保护

节点保护与链路保护

手工保护与自动（Auto）保护

其中，带宽保护与非带宽保护是根据用户配置，没有优先级之分。如果需要选择，则首先选择满足用户带宽配置需求的Bypass CR-LSP，如果同时满足用户带宽配置需求，则节点保护优于链路保护，手工保护优于自动保护。

 Page16 

一般来讲，节点保护可以同时保护被保护节点和 PLR 与被保护节点之间的链路，它看起来更优一些。如果可能的话，用户会更希望部署节点保护。华为公司提供了灵活的保护方式，在节点保护不能工作的时候，华为公司的设备支持保护方式自动降级为链路保护，当节点保护再次生效时，节点保护将重新起作用。

Bypass LSP的建立可以有两种方式，一种是手工方式，一种是自动方式;

- 手工 Bypass LSP是当一个没有快速重路由属性的隧道被指定保护一个物理接口以后，它所对应的 LSP就成为 Bypass LSP。手工 Bypass LSP的建立是通过在 PLR手工配置触发的。它的配置与普通 LSP基本没有分别，只是不能配置快速重路由属性。也就是说，Bypass LSP不能同时是主 LSP，LSP不能被嵌套保护。
- 自动 Bypass LSP是对手工方式的配置简化，当主 LSP需要被FRR保护时，PLR可以选择或自动建立一条 Bypass LSP，用来保护这个主 LSP，这种方式就叫自动 Bypass。自动 Bypass可以保护多个主 LSP，只要它可以满足这些主 LSP的要求。

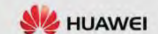


MPLS TE FRR的部署原则

TE FRR是MPLS TE中的一种局部性保护机制。在配置Bypass CR-LSP时，应该规划好它所保护的链路或节点，并确保该Bypass CR-LSP不会经过它所保护的链路或节点，否则不能真正起到保护作用。

FRR不支持多点故障。即，如果发生了FRR切换，数据从主CR-LSP切换到Bypass CR-LSP，在数据通过Bypass CR-LSP转发期间，Bypass CR-LSP的状态必须始终保持UP。一旦Bypass CR-LSP在此期间出现故障，被保护的数据将不能通过MPLS转发，从而可能出现流量中断，FRR功能失效。

FRR的Bypass隧道需要预先建立，这需要占用额外的带宽。在网络带宽余量不多的情况下，只能对关键的链路或节点进行FRR保护。

 Page17 

MPLS TE FRR技术不能解决作为隧道起始点和终结点的PE设备的故障。一旦PE节点发生故障，只能通过端到端的路由收敛、LSP收敛来恢复业务，其业务收敛时间与MPLS VPN内部路由的数量、承载网的跳数密切相关。在典型组网中一般在5s左右，无法达到节点故障端到端业务收敛小于1s的要求。


目 录
IP FRR技术
MPLS TE FRR技术
LDP FRR技术
VPN FRR技术
Copyright © 2012 Huawei Technologies Co., Ltd. All rights reserved.
Page18
HUAWEI

LDP FRR技术特点

不需要采用复杂的MPLS TE技术，设备开销小

本地化实现，无需相邻设备配合支持

多个节点分布式处理，备份端口可同时实现链路保护、节点保 护和路径保护

可用于对任意IP/MPLS流量的保护

Page19



MPLS TE FRR技术也有一些缺点，虽然出现了较长时间，但始终不能大规模部署。具体表现在：

- 1) 依赖于复杂的MPLS TE技术，设备开销大；
- 2) 备份LSP需手工显式的指定，配置工作量大；
- 3) 为进行链路、节点和路径保护，需要分别建立备份LSP，带来不必要的开销；
- 4) 备份LSP也存在故障可能，没有保护机制，当它失效时不能进行快速重路由；
- 5)要求备份LSP不能经过被保护的链路、节点，要求过于严格，有时候即使目的地可达，仍不能建立备份LSP。

为了克服上述缺点，需要一种设备开销小、配置简单、自适应网络变化的全新的技术方案，LDP FRR技术应运而生。

LDP FRR技术对LDP的要求

LDP FRR对LDP协议的要求：下游自主的标签分发+有序的标签控制+自由的标签保持。

在自由的标签保持方式下，LSR可以从任何相邻LSR收到对于FEC的标签映射消息，不论发送这一消息的相邻LSR是否是它所通告的特定FEC（标签映射中FEC所对应路由的下一跳），LSR对于它收到的所有标签映射都加以保留，其中，只有从FEC对应路由的下一跳发送来的标签映射会生成标签转发表（以下称为主标签转发表）。如果为其它标签映射也生成标签转发表，并作为主标签转发表的备份，相当于建立了备份LSP，LSR就可以快速对链路变化作出响应。

　Page20

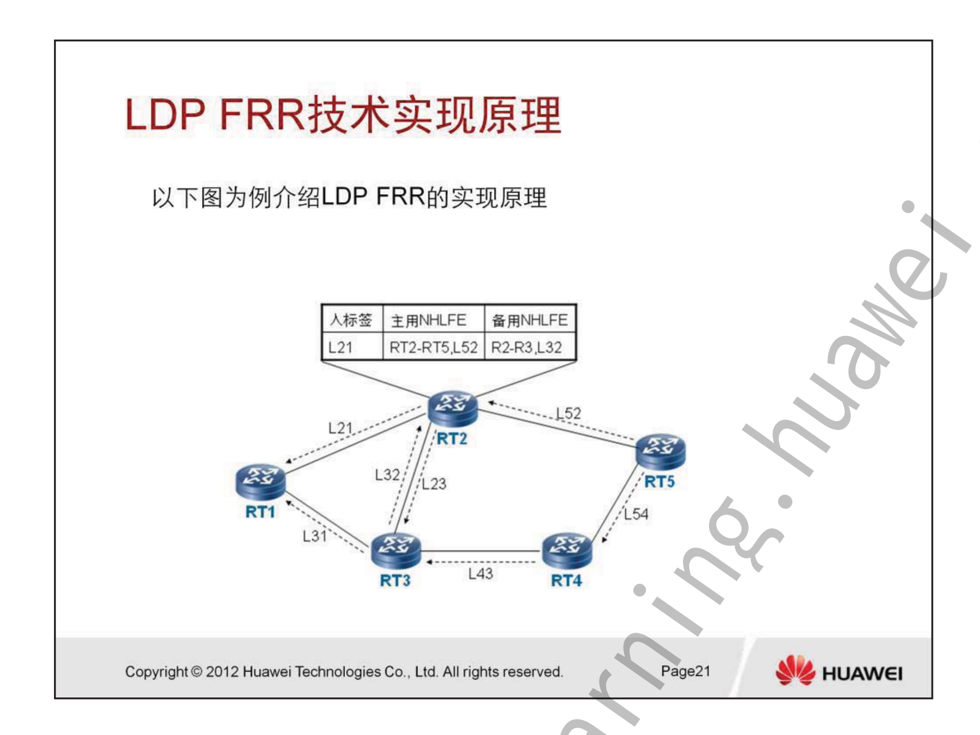


下游自主的标签分发（DU）：主动向上游LSR发送标签映射。

有序的标签控制：只有收到它的下游返回的标签映射消息后才向其上游发送标签映射消息。

自由的标签保持：保存所有邻居发送过来的标签。

DU+有序+自由的方式是当前的主流。



网络中运行LDP协议，其工作方式为DU（下游自主）标签分发+有序的标签控制+自由的标签保持。

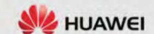
R1到达R5有多个路径，R5向上游发起多标签映射消息，最终，R2和R3分别给R1分配了到达R5的标签，其中，R2分配的标签主用，R3分配的标签可作为备用。

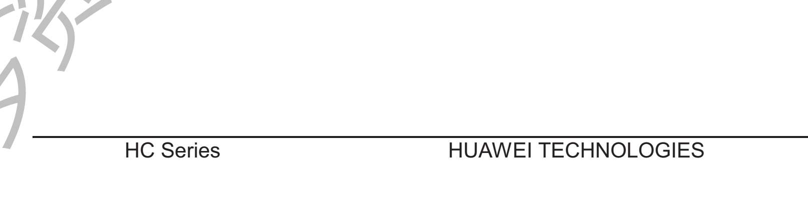
LDP FRR技术实现原理

1、指定LSR的一个设备端口作为另外一个设备端口的备份端口。这两个端口既可以是物理端口，也可以是逻辑端口，前提是LDP运行在这个端口上。

2、设备维护标签转发表，在未实施端口备份时，一个标签转发表仅有一个下一跳及标签，其中的标签是FEC的路由下一跳所连接LDP对等体为FEC分配的标签。在实施端口备份后，若某个标签转发表的下一跳是被保护的端口，为这个表项增加一个下一跳及标签，其中的标签是备份下一跳连接的LDP对等体为FEC分配的标签。如图所示，RT2上针对FEC（到达RT5的报文）生成两个NHLFE。

 Page22





LDP FRR技术实现原理（续）

3、设备维护每个端口的工作状态（正常/失效）。当检测到某个端口不能正常工作时 （如物理链路失效或人工操作将端口关闭），立即更新其状态。在报文转发过程中，查找标签转发表可以获得报文的下一跳端口，检查到其状态为失效，则倒换到备份的端口，并设置对应的标签，发送报文。

4、报文到达下一跳，由于标签是它自己分配的，这个下一跳上一定有对应的标签转发表，从而可以继续转发报文到目的地。

 Page23



由于是有序的方式，所以到达下一跳后可以继续转发报文。


目 录
IP FRR技术
MPLS TE FRR技术
LDP FRR技术
VPN FRR技术
Copyright © 2012 Huawei Technologies Co., Ltd. All rights reserved.
Page24
HUAWEI

VPN FRR技术概述

为了达到相邻节点业务倒换小于50ms、端到端业务收敛小于1s的要求，MPLS TE FRR技术、IGP路由快速收敛技术都应运而生，但是它们都无法解决在CE双归PE的网络中，PE设备节点故障时的端到端业务快速收敛的问题。

VPN FRR利用基于VPN的私网路由快速切换技术，通过预先在远端PE中设置指向主用PE和备用PE的主备用转发项，并结合PE故障快速探测，旨在解决CE双归PE的MPLS VPN网络中，PE节点故障导致的端到端业务收敛时间长（大于1s）的问题。VPN FRR简单可靠，部署方便，而且除了PE之间的故障快速检测机制之外，不依赖于周边设备的配合。

Page25



PE故障快速探测如：BFD、MPLS OAM等。

VPN FRR技术特点

实现在PE节点故障情况下，端到端业务收敛时间小于1s

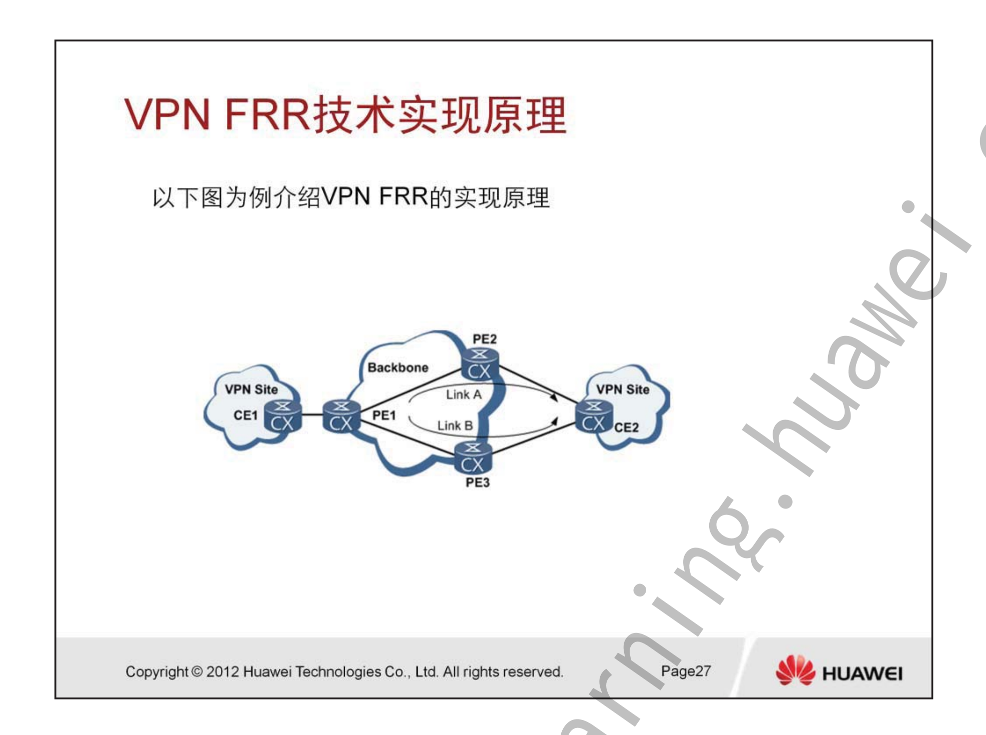
故障恢复时间与私网路由的规模无关

除了PE之间的故障快速检测机制之外，不依赖于周边设备的配合。

简单可靠，部署方便

 Page26



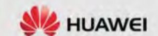


为了提高网络的可靠性部署CE双归PE之外，一般的，还会在PE2和PE3上部署VRRP协议，当作为VRRP主设备的PE2出现故障时，PE3成为新的VRRP主设备，并发布免费ARP报文，吸引从CE2访问CE1的流量从PE3上传；对于CE2访问CE1的流量，则利用VPN FRR技术，从PE2/PE3快速重路由到PE1，再由PE1下发给CE1，这个过程与VRRP的状态切换无关。

VPN FRR技术实现原理（续）

VPN FRR技术对传统技术进行了改进：支持PE1设备根据匹配策略选择符合条件的VPNv4路由；对于这些路由，除了优选的PE2发布的路由信息，次优的PE3发布的路由信息也同样填写在转发项中。

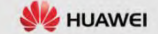
当PE2节点故障时，PE1通过BFD、MPLS OAM等技术感知到PE1与PE2之间的外层隧道不可用，便将LSP隧道状态表中的对应标志设置为不可用并下刷到转发引擎中，转发引擎命中一个转发项之后，检查该转发项对应的LSP隧道状态，如果为不可用，则使用本转发项中携带的次优路由的转发信息进行转发。这样，报文就会被打上PE3分配的内层标签，沿着PE1与PE3之间的外层LSP隧道交换到PE3，再转发给CE2，从而恢复CE1到CE2的业务，实现PE2节点故障情况下的端到端业务的快速收敛。

Page28


VPN FRR技术实现原理（续）

当L3VPN中承载了大量的路由时，按照传统的收敛技术，当远端PE出现故障时，所有这些VPN路由都需要重新迭代到新的隧道上，端到端业务故障收敛的时间与VPN路由的数量相关，VPN路由数量越大，收敛时间越长。而对于VPN FRR技术，我们只需要检测并修改外层隧道的状态，无论转发流量命中的是哪条VPN路由，流量都会切换到VPN FRR的备份路径上，其收敛时间只取决于远端PE故障的检测并修改对应隧道状态的时间，而与VPN路由的数量无关。

 Page29





总 结

本课程主要介绍了一下内容：

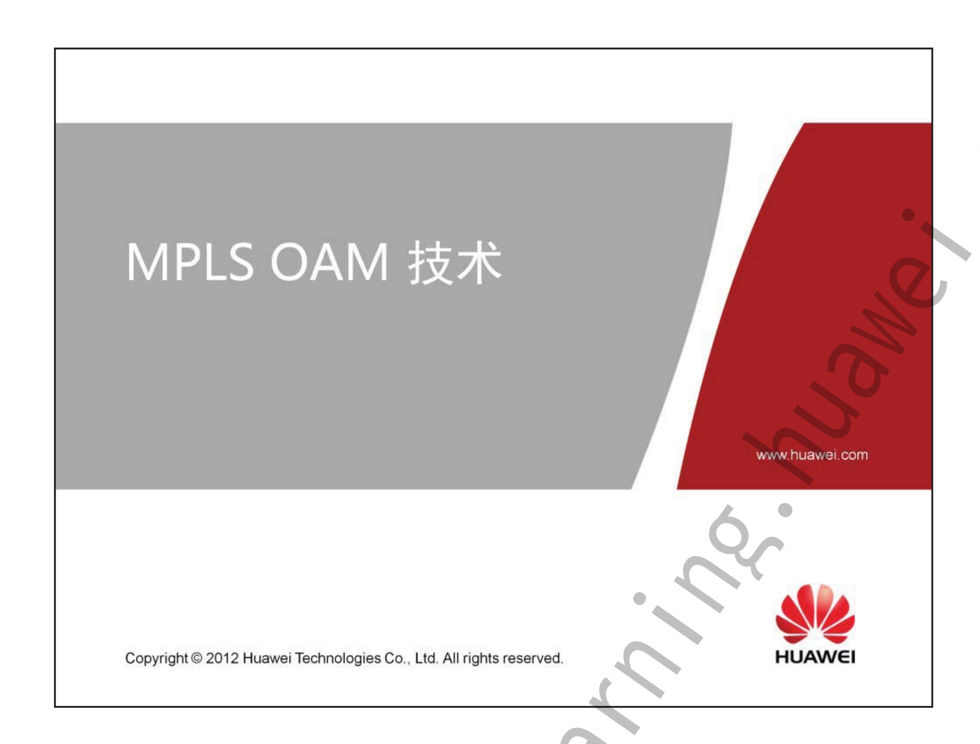
FRR技术的基本概念

目前主要应用的FRR技术的实现和基本原理：IP FRR，MPLS TE FRR，LDP FRR，VPN FRR

 Page30 

谢谢

www.huawei.com


MPLS OAM 技术
www.huawei.com
Copyright © 2012 Huawei Technologies Co., Ltd. All rights reserved.
HUAWEI



前 言

MPLS OAM技术为MPLS网络提供了一套缺陷检测的工具及缺陷纠正机制，通过MPLS OAM及保护倒换构件可以完成CR-LSP转发平面的检测功能，并在缺陷发生后的50ms内完成保护倒换，从而将缺陷所产生的影响减小到最低。

 Page1



OAM（Operation Administration & Maintenance）可以简化网络操作、随时检验网络性能、降低网络运行成本。部署有效的OAM机制对于公众电信网的运行非常重要，尤其是对于需要提供服务质量保障，即达到一定的性能和可用度要求的网络。



培训目标

学完本课程后，您应该能：

- MPLS OAM技术的基本概念
- MPLS OAM技术的基本原理

　Page2　HUAWEI


目 录
MPLS OAM技术概述
MPLS OAM检测技术
MPLS OAM保护倒换技术
Copyright © 2012 Huawei Technologies Co., Ltd. All rights reserved.
Page3
HUAWEI

MPLS OAM技术概述

承载MPLS的各种服务层，比如SDH都具有完善的OAM机制，问题在于MPLS可以在多种不同的服务层上传送（甚至LSP可以跨越由不同服务层组成的网络），而且它的用户层也是多种多样，如IP、FR、ATM、Ethernet等等，为了在MPLS的用户平面能确定LSP的连通性，MPLS层需要提供一种完全不依赖于任何用户层或物理层的OAM机制。

MPLS OAM实际上为MPLS用户层单独提供了一套检测机制，独立于其他网络层并为用户提供LSP的状态信息，为网络管理以及维护人员提供丰富的LSP诊断接口，为网络性能测量以及用户计费提供依据；MPLS OAM在提供检测工具的同时，还具备完善的保护倒换机制，能够在MPLS层发生缺陷后50ms内完成用户数据的倒换动作，使用户数据的损失减小的最低。

 Page4



MPLS作为可扩展的下一代网络的关键承载技术，提供具有QoS保障的多业务能力，并且，MPLS引入了一个新的网络层次，会存在由这个新的网络层引起的故障。因此，MPLS网络需要具备OAM能力。

承载MPLS的服务层（server-layer），例如SONET/SDH，以及利用MPLS的客户层（client-layer），例如IP、FR、ATM，都有各自的OAM机制。但MPLS网络层本身的故障不能完全通过其他层的OAM机制解决。并且，网络技术的分层要求，也需要MPLS具有自己独立的OAM机制，从而减少各层之间的依赖关系。

MPLS OAM需要实现以下功能：

- 提供按需查询和连续的检测，随时了解被监控的LSP是否存在缺陷。
- 发生缺陷后，能够探测到缺陷、分析、定位，通知网管系统，并根据缺陷的类型采取适当的行动。
    - 在链路出现缺陷或故障时迅速进行保护倒换，以便能根据与客户签订的SLA（Service Level Agreements）提供业务。并通过缩短业务中断的时间，减少维护时间和维护资源。



•在LSP迭代应用中，适当的处理可以抑制告警风暴。

•具备良好的兼容性，对于不支持OAM功能的标签交换节点LSN（Label Switching Node），丢弃OAM报文，但不影响用户流量，也不引发其他无关的动作。

•能够测量LSP的可用度和网络性能，为用户计费提供依据。


目 录
MPLS OAM技术概述
MPLS OAM检测技术
MPLS OAM保护倒换技术
Copyright © 2012 Huawei Technologies Co., Ltd. All rights reserved.
Page6
HUAWEI

MPLS OAM检测技术

MPLS OAM使用的报文分为三类：

连通性检测：包括两种类型的探测报文

- FFD（Fast Failure Detection）
- CV（Connectivity Verification）

前向缺陷通告FDI（Forward Defect Indication）

后向缺陷通告BDI（Backward Defect Indication）

MPLS OAM检测功能是指对TE LSP的连通性检测。MPLS OAM通过在被检测的TE LSP上周期性发送检测报文CV或FFD实现。

 Page7



FFD提供对P2P类型的LSP的检测机制。在LSP源端，FFD探测报文的产生速率可以调整，通常应用FFD来获得更快的检测速度。

OAM FFD报文同CV报文一样，也可以检测和诊断所有类型的LSP连通性缺陷，包括MPLS层以及MPLS层之下的缺陷。

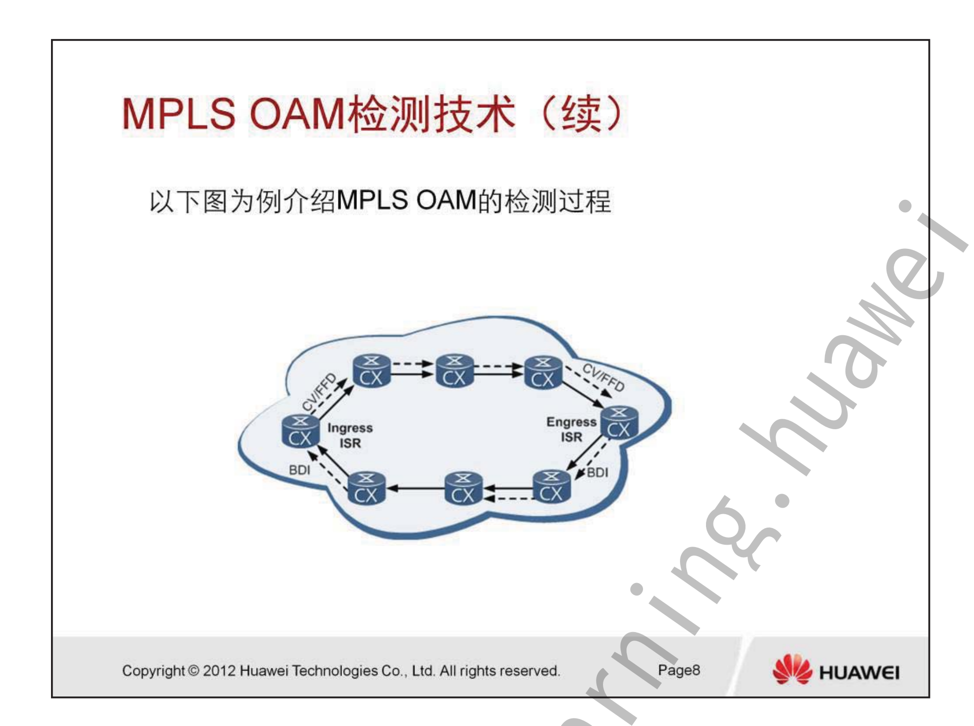
CV报文在LSP的源端LSR以每秒1个的速率产生，由LSP的宿端LSR接收。由于产生的速率较低且不可调整，通常用于普通的LSP可用度检测。

CV报文可用于检测和诊断所有类型的LSP连通性缺陷，包括MPLS层以及MPLS层之下的缺陷。

前向缺陷通告FDI报文是对CV或FFD检测到的缺陷进行的响应，由检测到缺陷的LSR产生，其主要目的是抑制受影响的客户LSP层产生告警。对于底层LSP：

- LSP的源端使用FDI报文将缺陷通告给LSP宿端。
- 启用自动协议时，LSP源端使用FDI报文通知宿端停止OAM检测。

下游LSP的宿端LSR检测到缺陷后，使用后向缺陷通告BDI报文，沿反向通道将缺陷告知上游的源端LSR。



BDI报文使用反向通道传送，反向通道可以是与被检测LSP具有相反入节点和出节点的LSP，也可以是连接被检测LSP的入节点和出节点的非MPLS路径。

具体来说，承载BDI报文的反向通道包括以下三种类型：

- 专用反向LSP：每条前向LSP有自己的反向LSP。这种方法相对稳定，但可能造成资源浪费。
- 共享反向LSP：多条前向LSP共用一条反向LSP，通过BDI携带的TTSI区分前向LSP。这种方法减少了资源浪费，但当多条前向LSP同时出现缺陷时，反向LSP可能发生拥塞。
- 非MPLS返回路径。这种方法存在安全隐患。例如，恶意用户可以制造一条BDI报文发给被检测LSP的源端，造成不必要的中断。

华为设备只支持前两种类型，即，只能使用LSP作为反向通道。

MPLS OAM检测技术（续）

MPLS OAM的检测过程：

入节点发送CV/FFD检测报文，报文通过被检测的LSP到达出节点。

出节点把接收到的报文类型、频率、TTSI等信息与本地记录的应该收到的值相比较，判断报文是否正确，并统计检测周期内收到的正确报文与错误报文的数量，从而对LSP的连通性进行监控。

CV报文的检测频率为固定值，FFD报文的检测周期为检测频率的三倍。

当出节点检测到LSP缺陷后，分析缺陷类型，通过反向通道将携带缺陷信息的BDI报文发送到入节点，从而使入节点及时获知缺陷状态。如果正确配置了保护组，还会触发相应的保护倒换。

 Page9



LSP源宿连接标识TTSI（Trail Termination Source Identifier）用于在网络中唯一标识一条LSP。

检测到的缺陷分为三类：非MPLS 层缺陷、MPLS 层缺陷、其他缺陷。

- 非MPLS 层缺陷：
  - dServer：服务层缺陷。任何来自MPLS 网络下层的非MPLS 层技术的服务层缺陷。
  - 承载MPLS 的下层网络可能会有其自身的保护及缺陷检测机制，当一条LSP某处出现底层缺陷后，距离该故障最近的LSR（下游方向）应能将该缺陷通告给Egress节点。对于发生的底层缺陷不应触发倒换动作，仅上报网管，但可以通过适当的方式通知Ingress 节点（BDI报文）。
  - dPeerME：对等实体缺陷。任何来自MPLS 子网外对等维护实体的非MPLS层技术的服务缺陷。
- MPLS 层缺陷：
  - dLOCV：连通性校验丢失缺陷。

- 在任意3 个连续的CV/FFD 发送周期内没有接收到相应的CV/FFD 报文即认为发生该缺陷。
- dTTSI_Mismatch：TTSI 失配缺陷。
- 在任意3 个连续的CV/FFD 发送周期内没有接收到带有正确TTSI 的CV/FFD报文即认为发生该缺陷。
- dTTSI_Mismerge：TTSI 错误合并缺陷。
- 在任意3 个连续的CV/FFD 发送周期内接收到既带有正确TTSI 又带有错误TTSI 的CV/FFD 报文，即认为发生该缺陷。
- dExcess：连通性检测报文超速缺陷。
- 在任意3 个连续的CV/FFD 发送周期内接收到超过（包括）5 个正确的CV/FFD 报文，即认为发生该缺陷。

其他缺陷：

- dUnknown：在MPLS 网络中出现未知缺陷。
- 这种缺陷可以自行定义用，比如Egress 检测到在同一条LSP 中既存在CV 报文又存在FFD 报文，类似这种协议没有规定的特殊缺陷可以用dUnknown 来标识。





保护倒换PS（Protection Switching）是为主Tunnel 预先建立相应的保护Tunnel（备用Tunnel）并为其分配带宽，主Tunnel 和备用Tunnel 构成一对保护组。当主Tunnel 发生缺陷时，数据流迅速倒换到备用Tunnel，减少由于LSP 失效造成的丢包或时延等问题，从而提高网络可靠性。保护倒换是端到端的保护。

配合MPLS OAM 的快速缺陷检测，可以使保护倒换达到毫秒级切换。

MPLS OAM保护倒换技术

1:1保护倒换：

1:1模式是在Tunnel的入节点和出节点间提供主备两条Tunnel

正常情况下，数据在主Tunnel传输。

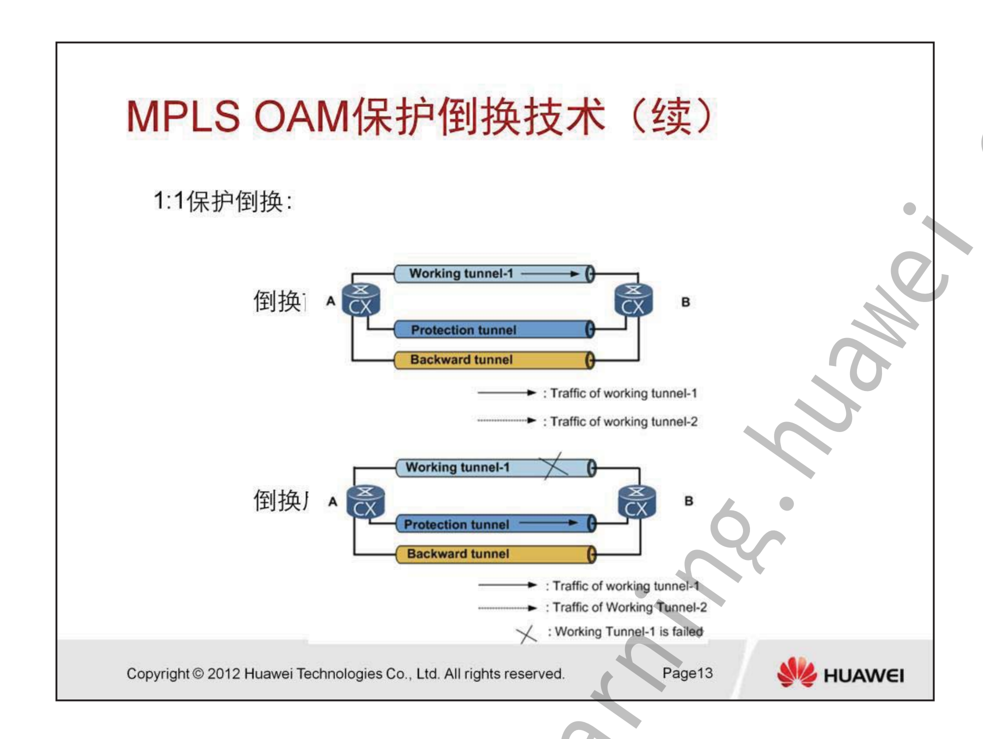
当入节点通过检测机制发现主Tunnel故障时，进行保护倒换，将数据切换到备用Tunnel上继续传输。

 Page12 

MPLS OAM保护倒换提供对整条LSP的保护，而不是对其中一段或某个节点的保护。

对选定的主LSP，备LSP的路由和带宽都是预留的。因此，保护倒换是一种完全指配的保护机制。为了保证在主LSP所有可能的失效情况下，保护都会有效实施，备LSP需要采用一条与主LSP完全不同的物理通道。





在1:1模式下，每条主LSP都有自己的备LSP。

正常情况下，数据通过主LSP传输，备LSP上没有主LSP的流量。

当宿端通过检测机制发现主LSP故障时，宿端先切换到备LSP上，再通过反向通道向源端发送BDI报文，通知Ingress将主LSP的流量旁路到备LSP上，从而完成1:1模式的保护切换。

反向通道用来向源端发送BDI报文，通知Ingress将主LSP的流量旁路到备LSP上，从而完成1:1模式的保护切换。

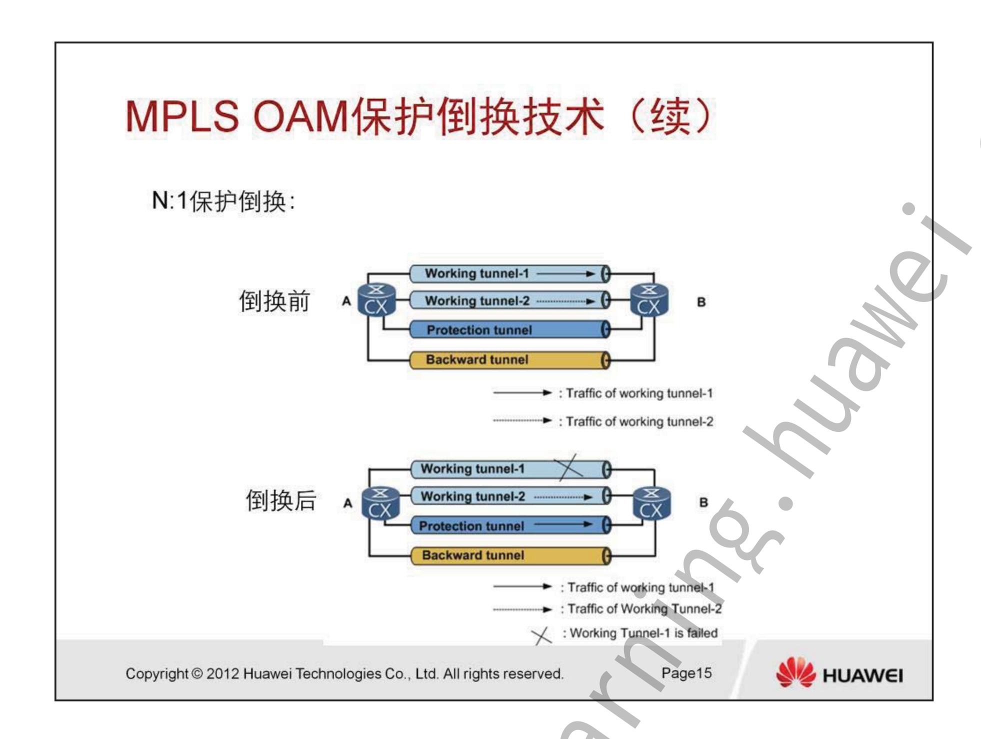
MPLS OAM保护倒换技术

N:1保护倒换：

N:1模式是将一条Tunnel作为多条主用Tunnel的备用Tunnel，当任何一条主用Tunnel故障时，都将数据倒换到共享的备用Tunnel上。这种模式主要是为了在网状拓扑结构的网络中节省带宽。

 Page14 





共享保护模式主要是为了在采用网状拓扑（mesh topology）的网络中节省带宽。

在共享保护模式下，使用一条LSP作为多条主LSP的备LSP，当任何一条主LSP故障时，都将数据倒换到共享的备LSP上。这种方式的触发机制和倒换机制与1:1模式类似。



总 结

本课程主要介绍了一下内容：

MPLS OAM技术的基本概念

MPLS OAM技术的检测原理

MPLS OAM技术的倒换原理





总 结

本课程主要介绍了一下内容：

MPLS OAM技术的基本概念

MPLS OAM技术的检测原理

MPLS OAM技术的倒换原理


谢谢
www.huawei.com
Copyright © 2012 Huawei Technologies Co., Ltd. All rights reserved.
HUAWEI


HA技术综合应用实例
www.huawei.com
Copyright © 2012 Huawei Technologies Co., Ltd. All rights reserved.
HUAWEI



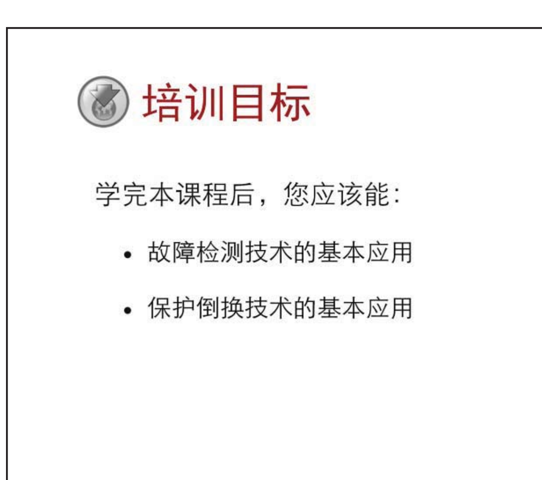
前 言

单独的故障检测技术或者保护倒换技术都不能完全实现业务的快速检测和切换，每一种故障检测技术和保护倒换技术都有自己的应用场景，本章以典型的IP承载网架构为例，从网络中的故障位置入手阐述各种可靠性技术在承载网上的综合部署。

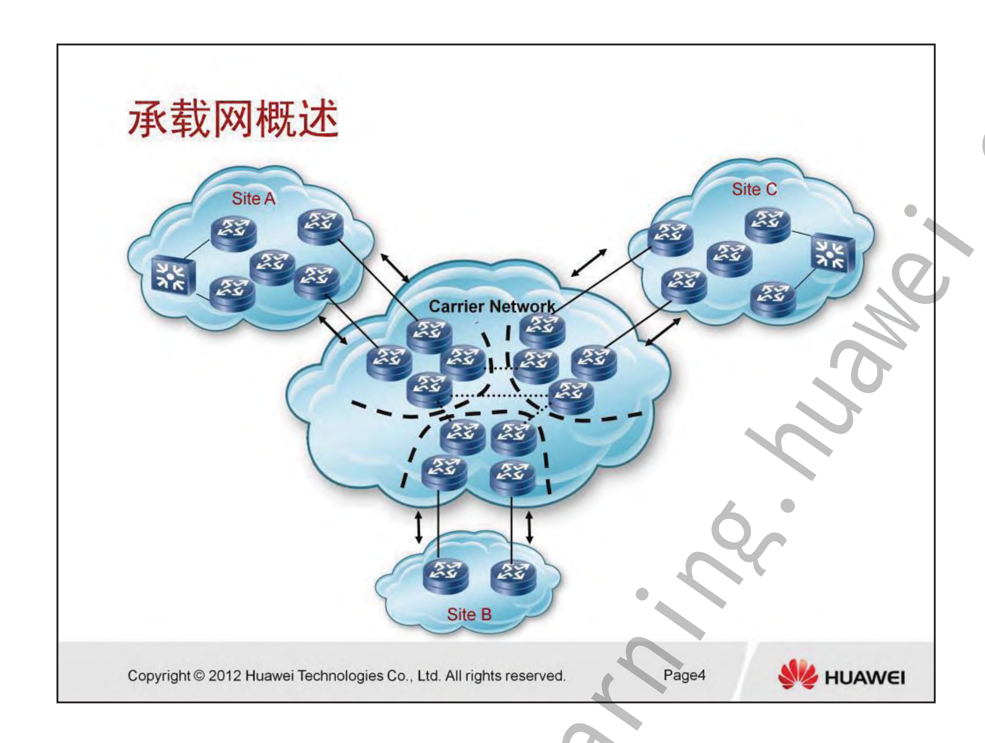
 Page1






培训目标
学完本课程后，您应该能：
• 故障检测技术的基本应用
• 保护倒换技术的基本应用
Copyright © 2012 Huawei Technologies Co., Ltd. All rights reserved.
Page2
HUAWEI


目 录
承载网概述
IP FRR应用
LDP/TE FRR应用
VPN FRR应用
Copyright © 2012 Huawei Technologies Co., Ltd. All rights reserved.
Page3
HUAWEI



传统IP网络仅承载Internet业务，采用非面向连接、尽力而为的服务模式，自愈时间长，安全性较差，不能满足承载电信级移动话音、视频业务的要求。如果要使用IP网络作为承载网，除了要对IP地址规划、IGP路由协议部署和MPLS部署进行精心设计以外，还必须在稳定性等方面做专门的优化和改进。

承载网概述（续）

对于用户来说这个网络是透明的，用户也不关心网络的实现细节，但肯定会关注网络的性能，这关系到他们的切身体会，而用户所关注的网络性能中最为重要的就是可靠性。

为了实现端到端的网络可靠性，则对于任何地方的单点故障都要有相应的HA技术作为可靠性的保证。

　Page5





目 录

 Page6



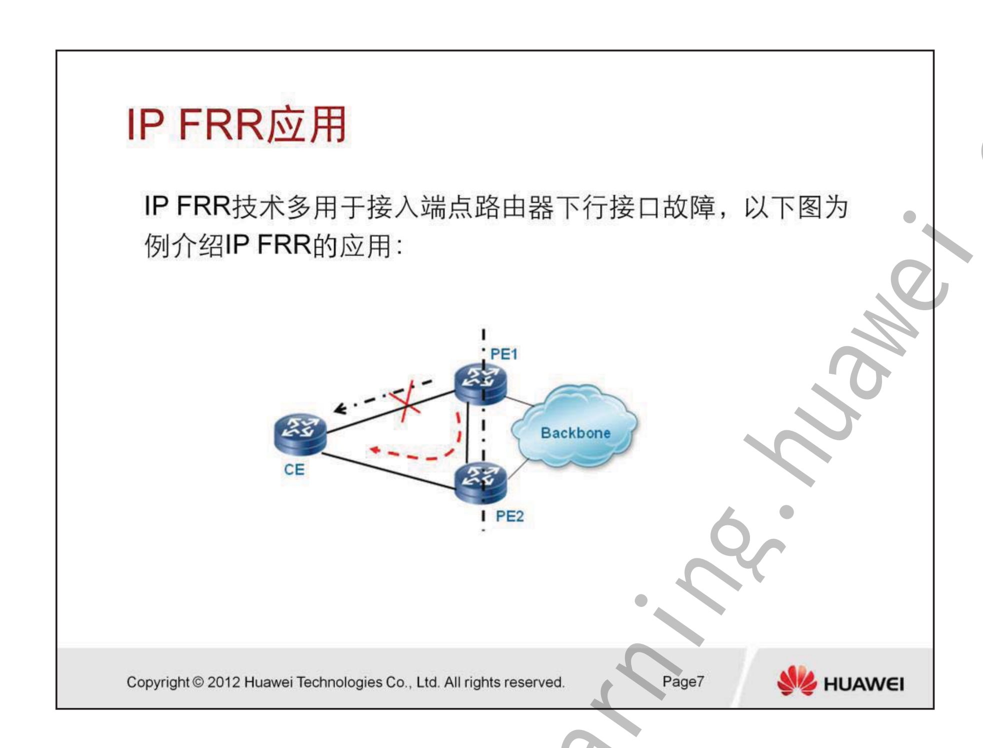
在转发模块中建立一张端口状态表，保存设备中每个端口的工作状态。当检测到端口不能正常工作时（如物理链路失效或人工操作将端口关闭），立即更新端口状态表。

同时，在报文转发过程中，如果查转发表得到的表项含有负载分担项（既有多个下一跳），在按照某种规则选定一个下一跳后，在端口状态表中检查这个下一跳出端口的状态，如果状态为失效，则使用另一个下一跳进行尝试，直至遍历全部负载分担项为止。

在检测到最后一个负载分担项时可以不再检查出端口的状态，而直接使用这个下一跳发送报文。

由于检测并更新端口状态的动作比路由收敛的动作快的多，所以本技术可以使重路由功能快速生效，并充分利用转发表中的多个负载分担项实现高可靠性的数据转发。

增强的IP FRR技术支持非等价负载分担的下一跳，IP路由有一个主用下一跳，由IGP计算决定，并手工配置一个备用端口（下一跳），在故障时快速切换。



通常还使用BFD技术来快速检测链路的故障，配合IP FRR做倒换。即在CE至PE1直接部署BFD。

IP FRR应用（续）

如上图所示，去往CE的流量经过PE-1(主用PE)转发，如果PE-1到CE的链路出现故障，此时我们使用IP FRR切换到PE-1到PE-2的链路。FRR的原理归根到底都是转发层面保持备份的路径，以做到快速切换。同样此处使用IP FRR时，PE-1到达CE的路径有两条，一是通过直连路径，另外一条就是通过PE-1到达PE-2再转发到CE。因为一般CE接入都会使用L3VPN，因此此处的IP FRR也是使用在私网下。这样我们PE-1和PE-2之间需要建立私网的邻居，可以在PE-1上实现到达CE有主备路径。

Page8



在路由收敛前，通过备份下一跳将流量切换至备份下一跳指定的链路，IP FRR起作用。在路由收敛后，按照路由选择的新的链路转发，IP FRR的接口备份使命结束。可见，备份下一跳的作用是填补了路由收敛的时间间隙，通过将流量快速切换到备份下一跳的备份链路，保证业务不中断。

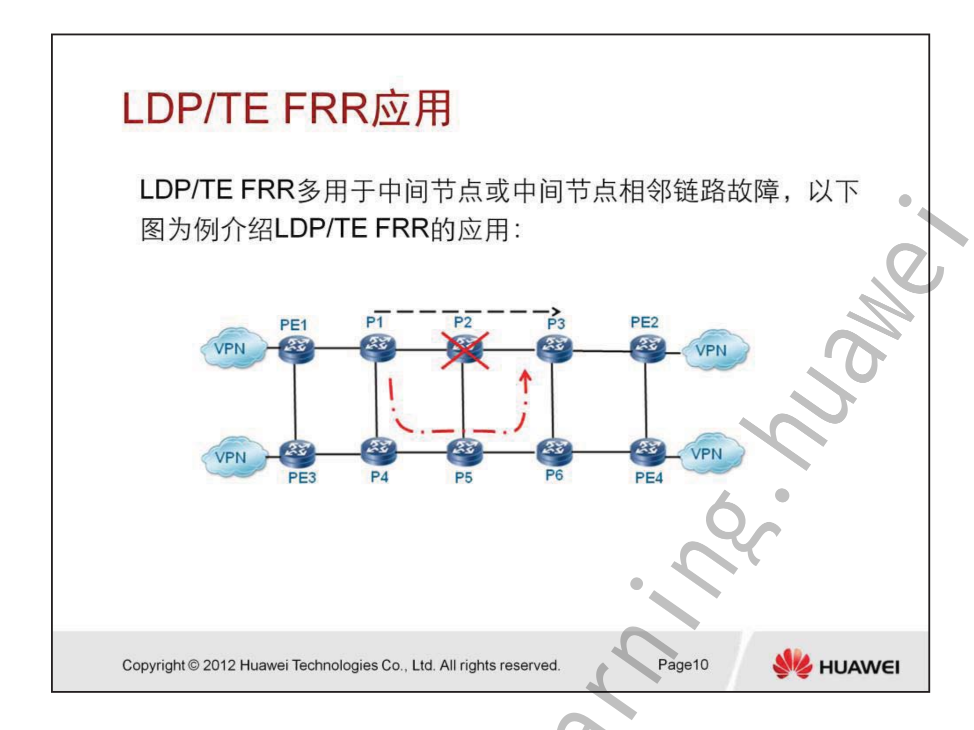
IP FRR针对IP网络路由而设计，分为公网IP FRR和私网IP FRR：

- 公网IP FRR：用于保护公网路由器。
- 私网IP FRR：用于保护CE路由器。



LDP FRR

- 传统的IP FRR无法有效保护MPLS网络中的流量，NE80E/40E提供LDP FRR功能，为MPLS网络提供端口级的保护方案。
- 与IGP快速收敛相比，LDP FRR事先计算备份端口，省去了故障发生后的路由计算时间以及LSP重新建立的时间，加快了保护倒换的速度。
- LDP工作在下游自主标签分发（DU）、有序标签控制（Ordered）以及自由标签保持方式（Liberal），所以LSR会保存所有收到的标签映射，但只有从FEC对应路由的下一跳发送来的标签映射会生成标签转发表。
- 利用这一特点，如果为Liberal标签映射也生成标签转发表，就相当于建立了备份LSP。
- 网络运行正常时，使用正常的LSP转发，如果正常LSP的出接口Down，使用备份LSP转发，这样可以在网络收敛之前的短时间内保证流量不中断。



MPLS TE FRR是现有的解决故障快速倒换的常用的技术之一，它的基本思路是在两个PE设备之间建立端到端的TE隧道，并且为需要保护的主用LSP事先建立好备用LSP，当设备检测到主用LSP不可用时（节点故障或者链路故障），将流量倒换到备用LSP上，从而实现业务的快速倒换。

从MPLS TE FRR技术的原理看，对于作为TE隧道起始点和终结点的两个PE设备之间的链路故障和节点故障，MPLS TE FRR能够实现快速的业务倒换。

但是这种技术不能解决作为隧道起始点和终结点的PE设备的故障。一旦PE节点发生故障，只能通过端到端的路由收敛、LSP收敛来恢复业务，其业务收敛时间与MPLS VPN内部路由的数量、承载网的跳数密切相关。在典型组网中一般在5s左右，无法达到节点故障端到端业务收敛小于1s的要求。

LDP/TE FRR应用（续）

如上图，承载网使用LDP作为公网隧道，在P路由器之间启用TE保障Qos。这种部署增强了全网的Qos能力，同时也降低了更换PE设备带来的TE部署的困难。在没有传输的情况下，如果故障发生在P1到P2之间或者P2故障，那么在P1本地开始TE FRR切换，保证切换时间在50ms之内，对LDP的业务基本不产生影响。如果故障发生在PE1到P1的链路，或者P1故障，LDP FRR在PE1上切换，保证切换时间在50ms之内。

上述的情况，有一个前提是没有传输设备，因为如果中间有传输设备，并且传输中间的链路产生问题，那么我们的路由器并不会感知到光电信号的中断，因此无法产生切换，此时，我们必须要另外一套机制来检测传输链路，那就是BFD和OAM。

 Page11



故障检测：

- 链路保护直接使用链路层协议实现故障检测和通告，链路层发现故障的速度与链路类型直接相关。节点保护则使用链路协议检测链路故障，在链路没有故障检测功能的情况下，可以通过配置BFD机制检测被保护节点的故障。
- 对于节点保护，只保护被保护节点及其与PLR （Point of Local Repair本地修复节点：Bypass CR-LSP的入节点，必须在主CR-LSP的路径上，可以是主CR-LSP的入节点，但不能是主CR-LSP的出节点）之间的链路。对于被保护节点和MP （Merge Point：汇聚点。Bypass CR-LSP的出节点，必须在主CR-LSP的路径上，并且不能是主CR-LSP的入节点）之间的链路故障，PLR无法感知。

无论是检测到链路故障还是节点故障，最终都会导致PLR上的出接口被置为老化状态。出接口老化就会触发FRR的流量切换。





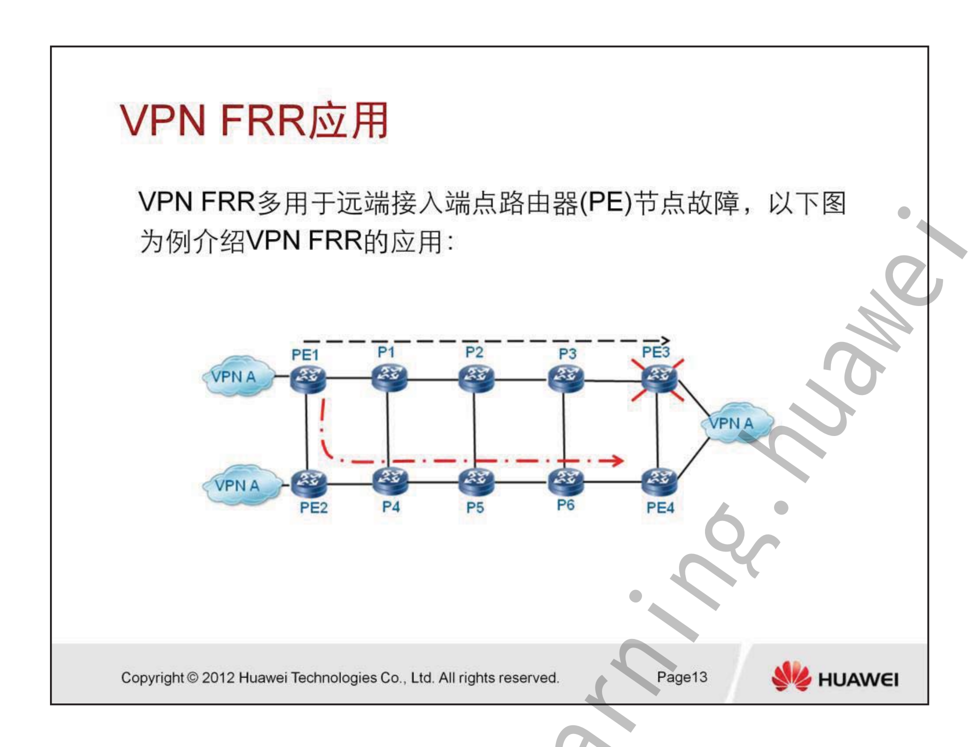
目 录

 Page12



为了达到相邻节点业务倒换小于50ms、端到端业务收敛小于1s的要求，MPLS TE FRR技术、IGP路由快速收敛技术都应运而生，但是它们都无法解决在CE双归PE的网络中，PE设备节点故障时的端到端业务快速收敛的问题。

VPN FRR利用基于VPN的私网路由快速切换技术，通过预先在远端PE中设置指向主用PE和备用PE的主备用转发项，并结合PE故障快速探测，旨在解决CE双归PE的MPLS VPN网络中，PE节点故障导致的端到端业务收敛时间长（大于1s）的问题。VPN FRR简单可靠，部署方便，而且除了PE之间的故障快速检测机制之外，不依赖于周边设备的配合。

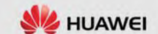


当L3 VPN中承载了大量的路由时，按照传统的收敛技术，当远端PE出现故障时，所有这些VPN路由都需要重新迭代到新的隧道上，端到端业务故障收敛的时间与VPN路由的数量相关，VPN路由数量越大，收敛时间越长。而对于VPN FRR技术，我们只需要检测并修改这些VPN路由迭代的外层公网隧道在转发引擎中的状态，无论转发流量命中的是哪条VPN路由，流量都会切换到VPN FRR的备份路径上，其收敛时间只取决于远端PE故障的检测并修改转发引擎中对应公网隧道状态的时间，而与VPN路由的数量无关。

VPN FRR应用（续）

承载网边缘接入路由器，如上图VPNA双归接入PE3、PE4，VPNA双归接入，在PE1上到达VPN A会有两个出口PE（PE3，PE4），这种情况下，我们可以让PE3和PE4互为备份，PE1上针对PE3和PE4的互相备份，我们称作VPN FRR。VPN FRR和所有的FRR相同，事先总会存在一个可用的备份路径，在主路径失效的情况下，能够做到快速切换。VPN FRR就是事先对于远端PE收到的私网路由保存两个下一跳（远端PE），这样我们可以确立一个为主，一个为备，选择主备PE的规则是由用户自己控制的。

 Page14



VPN FRR应用（续）

对于上图而言，PE1上，对于远端VPN A的路由，存在两个下一跳分别为PE3、PE4，此时PE1上面就可以选择PE3、PE4其中一个为主用的下一跳，另外一个为备用的下一跳。在没有配置VPN FRR的情况下，控制层面只会对转发层面下发一个主用的下一跳，当主用下一跳失效以后，备用下一跳再下发到转发层面，这个速度是比较慢的；在配置了VPN FRR以后，控制层面会把两个下一跳都下发到转发层面，这样在主用的下一跳失效以后，备用的下一跳在转发层面能够做到快速的使用，提高了切换速度，加上探测主下一跳即主PE失效的时间，使用BFD可以做到在100ms左右的时间内进行切换，实现了相当高的可靠性。

　Page15





总 结

本课程主要介绍了以下内容：

承载网的基本概念

各种可靠性技术（IP FRR，LDP/TE FRR，VPN FRR）的应用实例

 Page16

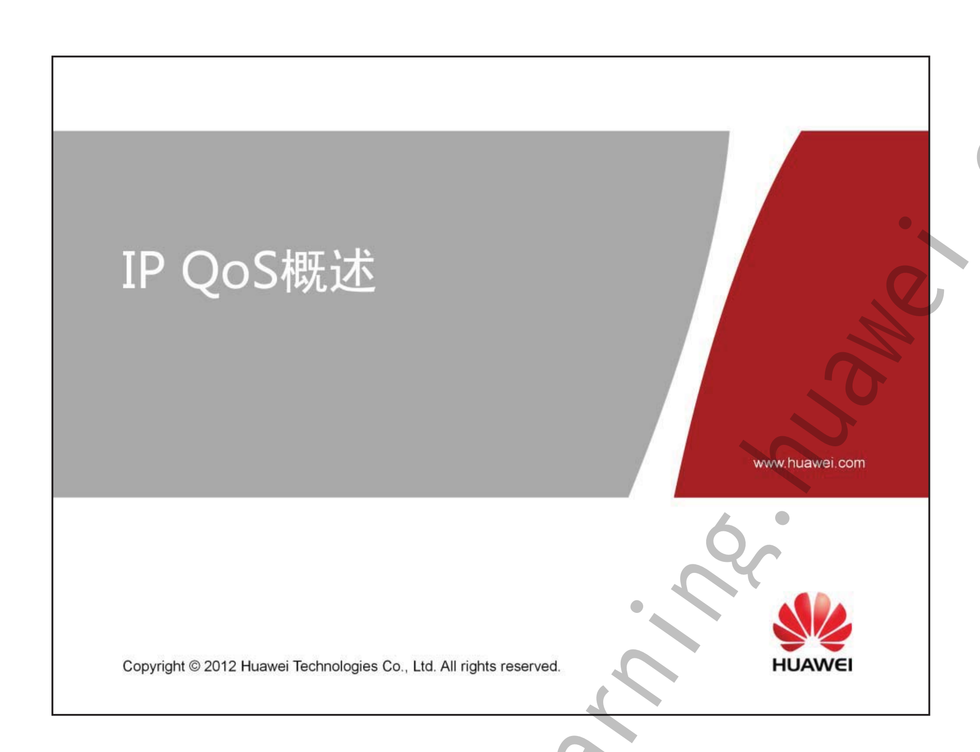


谢谢

www.huawei.com

Moudle 3

QoS


IP QoS概述
www.huawei.com
Copyright © 2012 Huawei Technologies Co., Ltd. All rights reserved.
HUAWEI



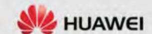
前 言

在传统的IP网络中，所有的报文都被无区别的等同对待，对报文传送的可靠性、传送延迟等性能不提供任何保证。

随着IP网络上新应用的不断出现，对IP网络的服务质量也提出了新的要求，例如VoIP等实时业务就对报文的传输延迟提出了较高要求，如果报文传送延时太长，用户将不能接受（相对而言，E-Mail和FTP业务对时间延迟并不敏感）。为了支持具有不同服务需求的语音、视频以及数据等业务，要求网络能够区分出不同的通信，进而为之提供相应的服务。传统IP网络的尽力服务不可能识别和区分出网络中的各种通信类别，而具备通信类别的区分能力正是为不同的通信提供不同服务的前提，所以说传统网络的尽力服务模式已不能满足应用的需要。

QoS技术致力于解决这个问题。

 Page1 HUAWEI




培训目标
学完本课程后，您应该能：
• 理解IP QoS的基本原理。
Copyright © 2012 Huawei Technologies Co., Ltd. All rights reserved.
Page2
HUAWEI

IP QoS的业务需求

传统的IP网络

主要承载数据业务，采用尽力传送（Best Effort）的方式，服务质量显得无关紧要。

当前的IP网络

近年来，随着以IP技术为核心的Internet的飞速发展，以及各种新业务的出现（VoIP、VPN、ERP等），IP网络已由一个单纯的数据网络转变为具有商业价值的承载网，因此IP网络必须为其所承载的每一类业务提供相应的服务质量。

 Page3 

在传统的IP网络中，所有的报文都被无区别的等同对待，每个路由器对所有的报文均采用先入先出（FIFO）的策略进行处理，它尽最大的努力（Best-Effort）将报文送到目的地，但对报文传送的可靠性、传送延迟等性能不提供任何保证。

随着IP网络上新应用的不断出现，对IP网络的服务质量也提出了新的要求，例如VoIP（Voice over IP，IP语音）等实时业务就对报文的传输延迟提出了较高要求，如果报文传送延时太长，将是用户所不能接受的。为了支持具有不同服务需求的语音、视频以及数据等业务，要求网络能够区分出不同的通信，进而为之提供相应的服务。传统IP网络的尽力服务不可能识别和区分出网络中的各种通信类别，传统网络的尽力服务模式已不能满足应用的需求。QoS（Quality of Service，服务质量）技术的出现便致力于解决这个问题。

QoS旨在针对各种应用的不同需求，为其提供不同的服务质量，例如：提供专用带宽、减少报文丢失率、降低报文传送时延及时延抖动等。

IP QoS的概念

QoS：Quality of Service，即服务质量

IP QoS 是指IP网络的一种能力，即在跨越多种底层网络技术（MP、FR、ATM、Ethernet、SDH、MPLS等）的IP网络上，为特定的业务提供其所需要的服务。服务质量包括：

- 传输的带宽
- 传输的时延和抖动
- 数据的丢包率

网络中存在资源竞争，就存在对服务质量的要求

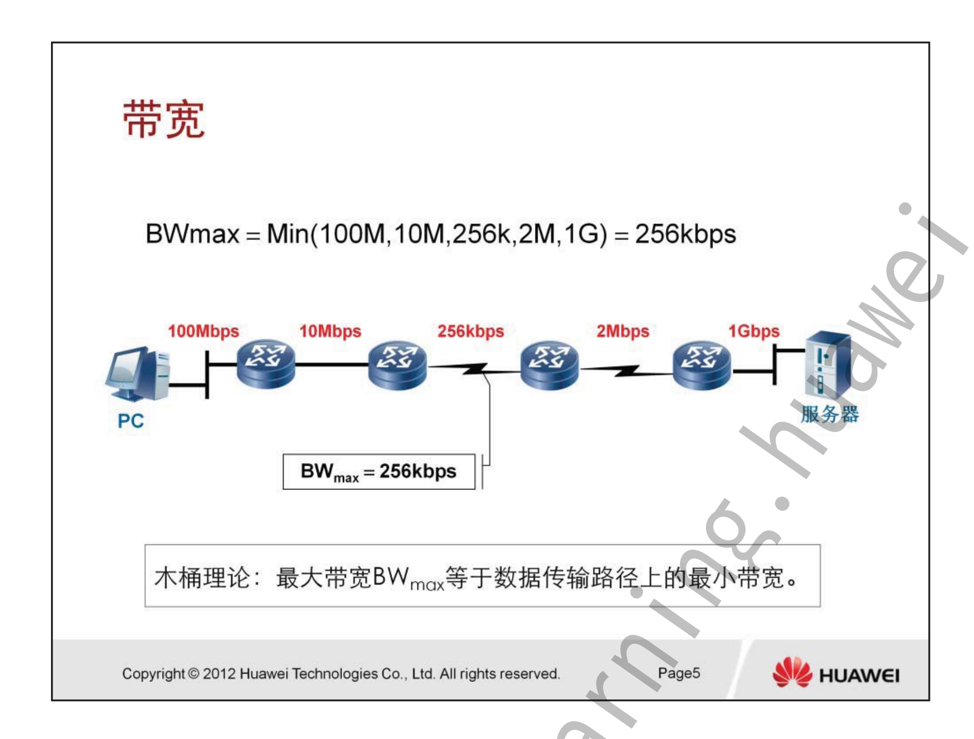
提高某类业务的服务质量同时也会损害其它业务的服务质量

 Page4 

QoS，顾名思义，Quality of Service，服务质量。对于网络业务来说，服务质量包括哪些方面呢？从传统意义上来讲，无非就是传输的带宽、传送的时延、数据的丢包率等，而提高服务质量无非也就是保证传输的带宽，降低传送的时延和时延抖动，降低数据的丢包率等。广义上讲，服务质量涉及网络应用的方方面面，只要是对网络应用有利的措施，其实都是在提高服务质量。因此，从这个意义上来说，防火墙、策略路由、快速转发等也都是提高网络业务服务质量的措施之一。

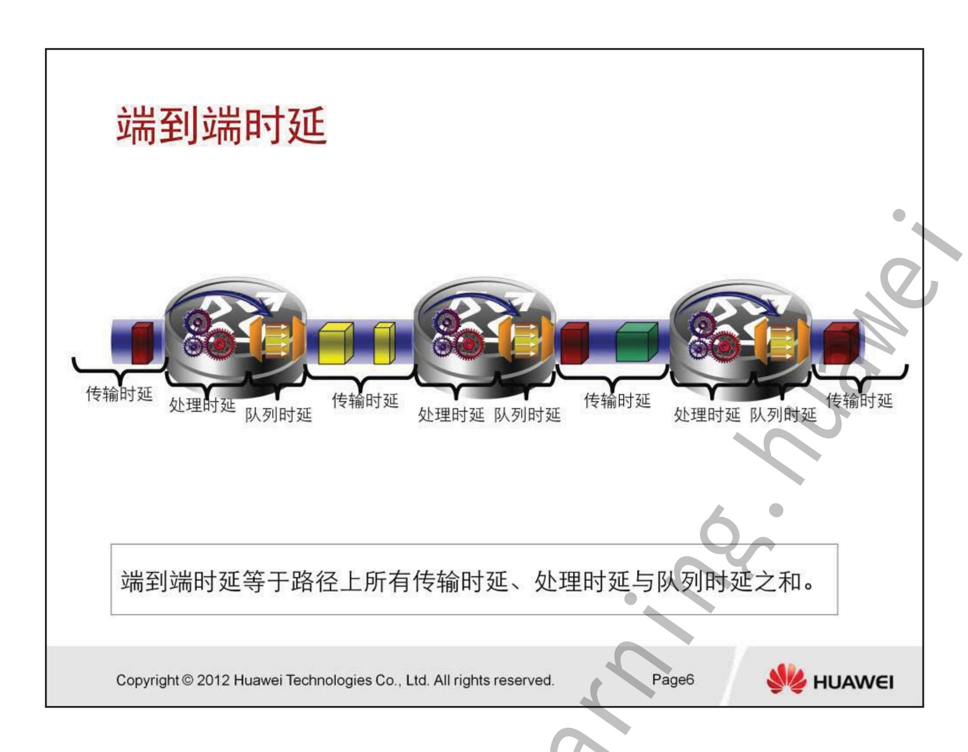
但是，服务质量又是相对网络业务而言的，在保证了某类业务的服务质量的同时，可能也在损害其它业务的服务质量。因为网络资源总是有限的，只要存在竞争网络资源的情况，就会出现服务质量的要求。比如，网络总带宽为100Mbps，而BT下载占用了90Mbps，其他业务就只能占用剩下的10Mbps。而如果限制BT下载占用的最大带宽为50Mbps，则也就提高了其他业务的服务质量，使其他业务能够占用最少50Mbps的带宽，但是这是在损害BT业务的服务质量为前提的。





带宽决定数据传输的速率，例如100Mbps的带宽意味着在理论上数据可以以100M比特每秒的速率进行传输。

整个传输途径的带宽取决于这条途径上的最小链路带宽。如图所示，尽管传输途径上的最大的一段链路带宽是1Gbps，但是数据从PC传到服务器，最大的传输速率只能是256kbps，因为传输的最大带宽是由传输路径上的最小链路带宽决定的。正是因为这样，带宽小的链路是影响传输速率的关键。



端到端的时延由传输时延、处理时延和队列时延组成。

传输时延又叫串行化时延，它的大小在很大程度上取决于带宽的大小。

处理时延是指路由器把数据包从入接口放到出接口队列需要的时间，它的大小跟路由器的处理性能有关。

队列时延指数据包在出口队列中停留的时间，它的大小跟队列中数据包的大小和数量、带宽以及队列机制有关。

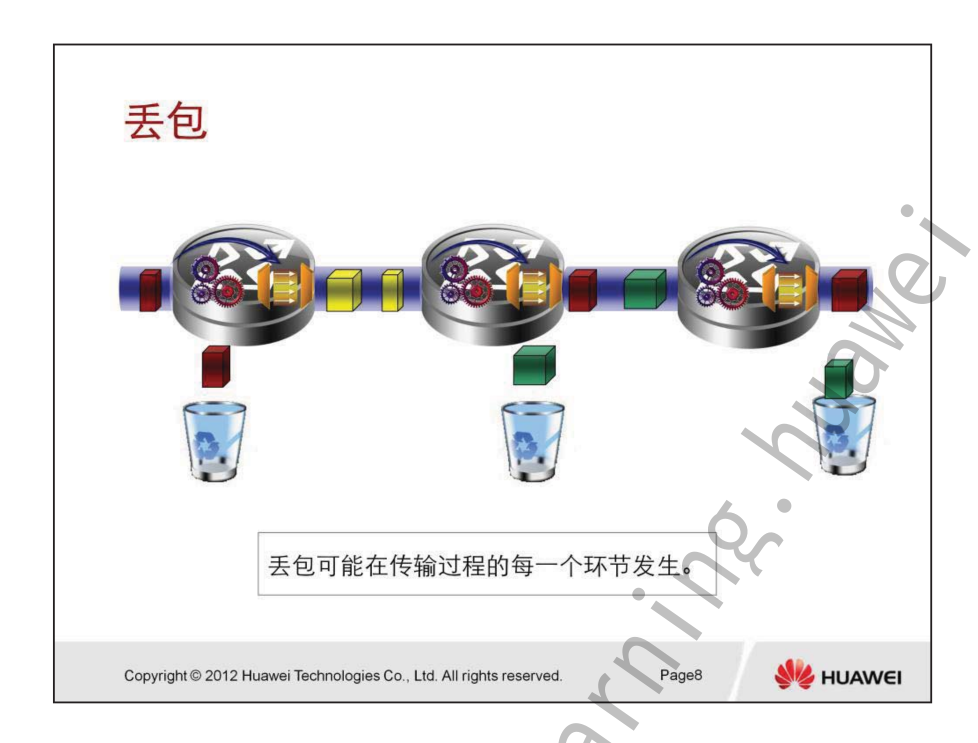
另一个与时延相关的概念是时延抖动，时延抖动是由于属于同一个流的数据包的端到端时延不相等造成的，一般来说，时延越小则时延抖动的范围越小。



抖动是由于属于同一个流的数据包的端到端时延不相等造成的。

图中源端等间隔发送数据包，因为每个数据包的端到端时延不一样导致这些包不能等间隔到达目的端，这种现象叫做抖动。

抖动的大小跟时延的大小直接相关，时延小则抖动的范围也小，时延大则可能的抖动范围也大。



丢包可能在所有环节上发生，比如：

路由器在收到数据包的时候因为CPU繁忙，没办法处理数据包，导致丢包；

在把数据包调度到队列的时候因为队列满而导致丢包；

数据包在链路上传输的时候因为种种原因（链路故障、冲突）等导致的丢包。

在很多时候，丢包一般是因为队列满造成的，在队列满的时候，一般采用尾丢弃（把最后到达的包丢弃）来进行丢包。

QoS服务模型

QoS 根据网络质量和用户需求，通过不同的服务模型为用户提供服务。通常QoS 提供以下三种服务模型：

- Best-Effort Service模型：是最简单的服务模型。应用程序可以在任何时候，发出任意数量的报文，而且不需要事先获得批准，也不需要通知网络。
- Integrated Service模型：是一个综合服务模型，它可以满足多种QoS需求。这种服务模型在发送报文前，需要向网络申请特定的服务。
- Differentiated Service模型：是通过设置报文头部的QoS 参数信息，来告知网络节点它的QoS需求。报文传播路径上的各个路由器都可以通过对报文头的分析来获知报文的服务需求类别。

 Page9



1、Best-Effor Servicet： 尽力而为的服务模型，Best-Effort是一个单一的服务模型，也是最简单的服务模型。应用程序可以在任何时候，发出任意数量的报文，而且不需要事先获得批准，也不需要通知网络。对Best-Effort服务，网络尽最大的可能性来发送报文，但对时延、可靠性等性能不提供任何保证。 Best-Effort服务是现在Internet的缺省服务模型，它适用于绝大多数网络应用，如FTP、E-Mail等，它通过先入先出（FIFO）队列来实现。

2、IntServ（Integrated Service）集成服务模型。 Intserv是一个综合服务模型，它可以满足多种QoS需求。这种服务模型在发送报文前，需要向网络申请特定的服务。这个请求是通过信令（signal）来完成的。应用程序首先通知网络它自己的流量参数和需要的特定服务质量请求，包括带宽、时延等，应用程序一般在收到网络的确认信息，即确认网络已经为这个应用程序的报文预留了资源后，才开始发送报文。同时应用程序发出的报文应该控制在流量参数描述的范围以内。

网络在收到应用程序的资源请求后，执行资源分配检查（Admission control），即基于应用程序的资源申请和网络现有

的资源情况，判断是否为应用程序分配资源。一旦网络确认为应用程序的报文分配了资源，则只要应用程序的报文控制在流量参数描述的范围内，网络将承诺满足应用程序的QoS需求。而网络将为每个流（flow，由两端的IP地址、端口号、协议号确定）维护一个状态，并基于这个状态执行报文的分类、流量监管（policing）、排队及其调度，来实现对应用程序的承诺。

在IntServ服务模型中，负责传送QoS请求的信令是RSVP（Resource Reservation Protocol，资源预留协议），它通知路由器应用程序的QoS需求。RSVP是在应用程序开始发送报文之前来为该应用申请网络资源的，所以是带外（out-bind）信令。

IntServ可以提供以下两种服务：

(1)、保证服务（Guaranteed service） 它提供保证的带宽和时延限制来满足应用程序的要求。如VoIP应用可以预留10M带宽和要求不超过1秒的时延。

(2)、负载控制服务（Controlled-Load service） 它保证即使在网络过载（overload）的情况下，能对报文提供近似于网络未过载类似的服务，即在网络拥塞的情况下，保证某些应用程序的报文低时延和高通过。

3、DiffServ（Differentiated Service）差分服务模型。

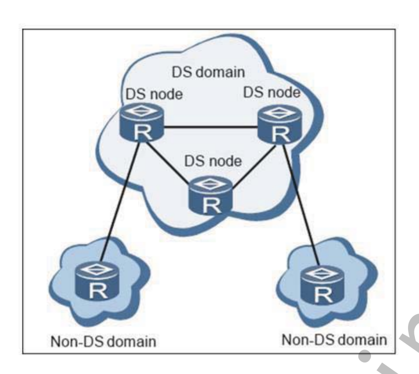
Differentiated Service 模型即区分服务模型，简称Diff-Serv。在采用Diff-Serv 模型的应用中，应用程序在发送报文前不必预先向网络提出资源申请，而是通过设置IP 报文头部的QoS 参数信息，来告知网络节点它的QoS 需求。报文传播路径上的各个路由器都可以通过对IP 报文头的分析来获知报文的服务需求类别。

在实施Diff-Serv 时，接入路由器需要对报文进行分类，并在IP 报文头部标记服务类别。下游的路由器只需简单地识别这些服务类别，并进行转发。因此，Diff-Serv 是一种基于报文流的QoS 解决方案。



IP 网络中的Diff-Serv 模型

一般来讲，在提供IP网络的QoS时，为了适应不同规模的网络，在IP骨干网往往需要采用DiffServ体系结构。Diff-Serv 网络结构示意图：



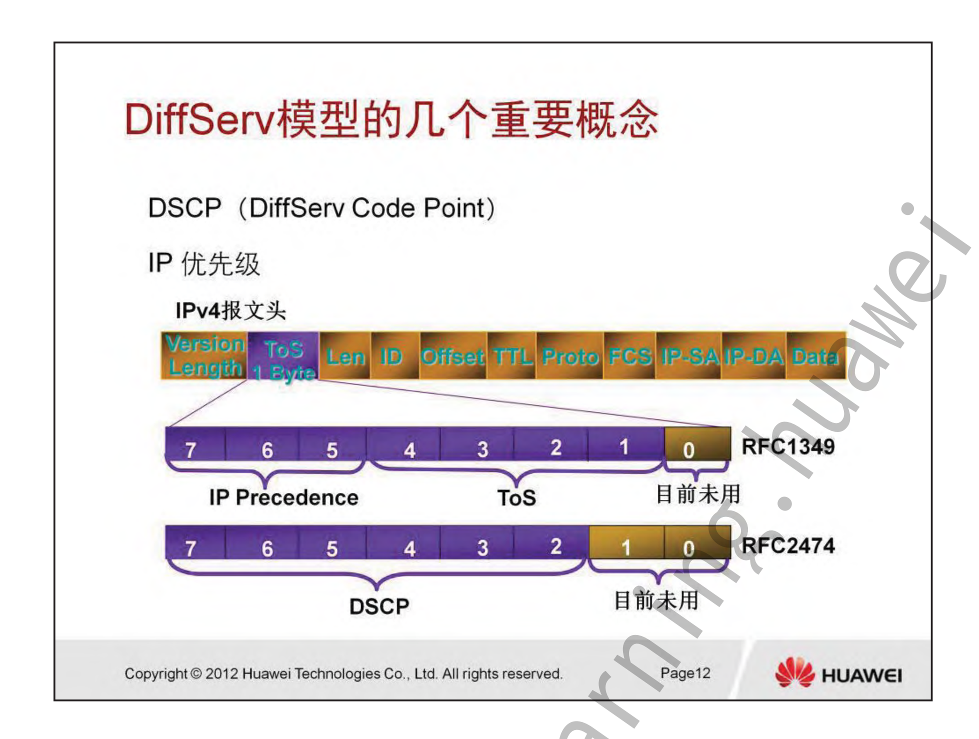
 Page11 

实现了Diff-Serv 功能的网络结点称为DS 节点。DS（DS Domain）域由一组采用相同的服务提供策略和实现了相同PHB（Per-Hop Behavior）集合的相连DS 节点组成，如上图 所示。

DS 节点分为两种：

1、DS 边界节点，用于将DS 域和非DS 域连接在一起。DS 边界节点需根据域间制定的流量控制协定TCA（Traffic Conditioning Agreement）进行流量控制并设置报文的DSCP（Differentiated Services CodePoint）值。

2、 DS 内部节点，用于在同一个DS 域中连接DS 边界节点和其他内部节点。DS 内部节点仅需基于DSCP 值进行简单的流分类以及对相应的流实施流量控制



在IP报文中有专门的字段进行QoS的标记，在IPV4中为ToS，IPv6中为TrafficClass。ToS字段用前6bit来标记DSCP，如果只用前3 bit 就为IP优先级。DSCP和IP优先级都是标记的标准。

IP优先级提供0-7共8种服务质量，6和7都保留所以常用的是0-5，每个数字都对应一个名称，比如0对应Routine，这样在更改数据包优先级等配置时，既可以用数字也可以用名称。

注意优先级中的数字本身没有实际的意义，标记为5的数据优先级不一定就比标记为0的高，只是一个分类标准而已。真正的操作是在配置上针对不同的优先级采取不同的措施，比如什么标识的数据包属于什么队列。

不管是IP优先级还是DSCP都是用自己的前3bit和二层的CoS值形成映射。

在二层用CoS字段进行标记，正常的以太网帧是没有标记的，但是在ISL的报头和802.1Q的Tag中都有3bit 用来定义服务级别，从0到7，不过只有0-5可用，6和7都保留。

DiffServ模型的几个重要概念（续）

PHB（Per-Hop Behaviors），PHB是DS节点作用于数据流的行为。网络管理员可以配置DSCP到PHB的映射关系。如果DS节点接收到一个报文，检查其DSCP，发现未定义到PHB的映射，则DS节点将选择采用缺省PHB（即Best-Effort，DSCP=000000）进行转发处理。每个DS节点必须支持该缺省PHB。

PHB的分类，IETF DiffServ工作组目前定义了四种PHB：

- Default PHB
- Class-Selector PHB
- Expedited Forwarding PHB
- Assured Forwarding PHB

 Page13



PHB（Per-Hop Behavior）是DS 节点作用于数据流的行为。网络管理员可以配置DSCP 到PHB 的映射关系。PHB是网络节点对报文调度、丢包、监管和整形的处理，每类PHB都对应一组DSCP；PHB只定义了一些外部可见的转发行为，没有指定特定的实现方式。

目前，IETF 定义了四种标准的PHB：类选择码CS（Class Selector），加速转发EF（Expedited Forwarding），确保转发AF（Assured Forwarding）和尽力而为BE（Best-Effort）。其中，BE 是缺省的PHB。

1、 CS PHB

CS 表示类选择码，代表的服务等级与在现有网络中使用的IP Precedence 相同。DSCP 取值为 “XXX000”，X 为0 或1。当X为全0 时，就是Default PHB。

2、 EF PHB

DSCP为 “101110” RFC2598；代表DiffServ网络中最高的服务质量，在有带宽确保的情况下，发包速度大于收包速度，适用于VoIP、虚拟租用线等实时业务；可通过优先队列、低时延队列或

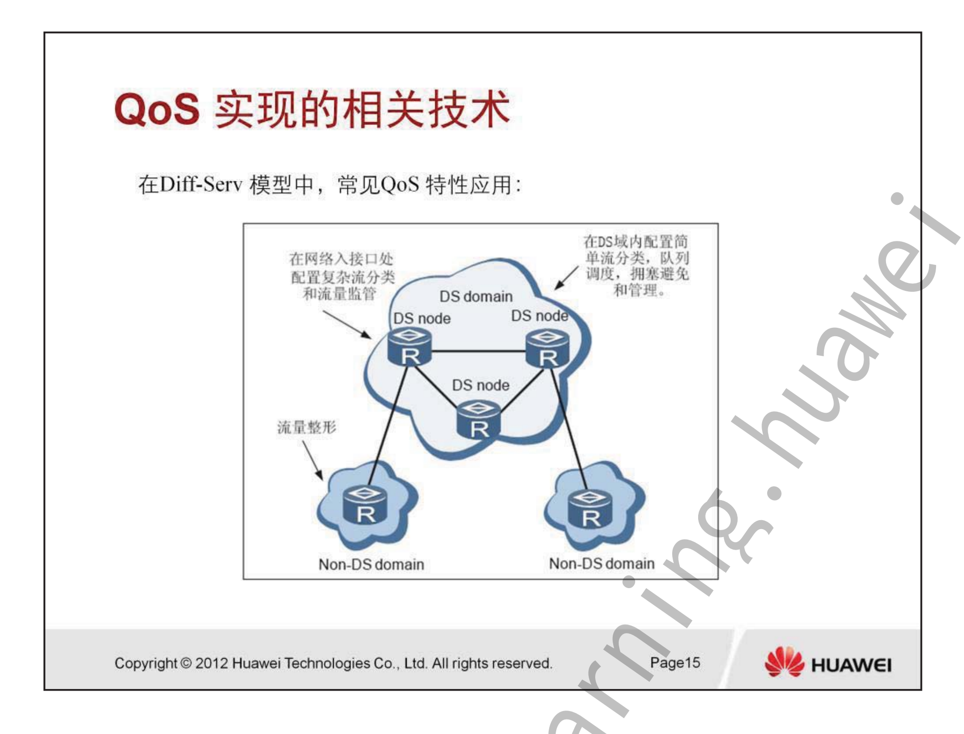
RTP实时队列等多种队列机制来实现加速转发被定义为这样的一种转发处理：从任何DS 节点发出的信息流速率在任何情况下必须获得等于或大于设定的速率。EF PHB 在DS 域内不能被重新标记。仅允许在边界节点重新标记EF PHB，并且要求新的DSCP 满足EF PHB 的特性。定义EF PHB 的目标是在DS 域内模拟一种虚拟租用线（Virtual Leased Line）的转发效果，提供一种低丢包率、低延迟、高带宽的转发服务。

3、 AF PHB

确保转发的推出是为了满足这样的需求。用户在与ISP 订购带宽服务时，允许业务量超出所订购的规格。对不超出所订购规格的流量要求确保转发的质量；对超出规格的流量将降低服务待遇继续转发，而不只是简单地被丢弃。当前定义了四类AF，即AF1、AF2、AF3、AF4。每一类AF 业务的分组又可以细分为三种不同的丢弃优先级。AF 编码点AFij 表示AF 类为i（1<=i<=4），丢弃优先级为j（1<=j<=3）。运营商在提供AF 服务时，为每类AF 分配不同的带宽资源。对AF PHB 的一个特别要求是：流量控制不能改变同一信息流中分组的顺序。比如，某一业务流中的不同分组归属同一AF 类，但在流量监管时被标记了的不同的丢弃优先级，此时，虽然不同分组的丢包概率不同，但是他们之间的相互顺序不能改变。这种机制特别适合于多媒体业务的传输。

4、BE PHB即传统的IP 分组投递服务，只关注可达性，其他方面不做任何要求。任何路由器必须支持BE PHB。





流分类、流量监管、流量整形、拥塞管理和拥塞避免是构造有区别地实施服务的基石。流分类是基础，它依据一定的匹配规则识别出报文，是有区别地实施服务的前提。而流量监管、流量整形、拥塞管理和拥塞避免从不同方面对网络流量及其分配的资源实施控制，是有区别地提供服务思想的具体体现。它们主要完成如下功能：

1、流分类：依据一定的匹配规则识别出对象。流分类是有区别地实施服务的前提。

2、流量监管：对进入路由器的特定流量的规格进行监管。当流量超出规格时，可以采取限制或惩罚措施，以保护运营商的商业利益和网络资源不受损害。

3、流量整形：一种主动调整流的输出速率的流控措施，通常是为了使流量适配下游路由器可供给的网络资源，避免不必要的报文丢弃和拥塞。

4、拥塞管理：网络拥塞时必须采取的解决资源竞争的措施。通常是将报文放入队列中缓存，并采取某种调度算法安排报文的转发次序。

5、拥塞避免：过度的拥塞会对网络资源造成损害。拥塞避免监督网络资源的使用情况，当发现拥塞有加剧的趋势时采取主动丢弃报文的策略，通过调整流量来解除网络的过载。



问 题

什么是QoS?

QoS包括哪些方面?

常见的QoS服务模型有哪些?

IPv4报文中，DSCP、ToS和IP Precedence的关系?

 Page17



答案:

- QoS：即服务质量。
- QoS包括：传输的带宽，传输的时延和抖动，数据的丢包率；
- 常见的QoS服务模型有Best-Effort、 IntServ(Integrated Service）、DiffServ(Differentiated Service)。
- DSCP、TOS和IP Precedence的区别：在IPV4报文头中，有1Byte的字段表示ToS，ToS字段的高6bit可以表示DSCP,高3bit可以表示IP Precedence。

谢谢
www.huawei.com


流量分类与标记
www.huawei.com
Copyright © 2012 Huawei Technologies Co., Ltd. All rights reserved.
HUAWEI



前 言

为了在Internet上针对不同的业务提供有差别的QoS服务质量，人们根据报文头中的某些字段记录QoS信息，从而让网络中的各设备根据此信息提供有差别的服务质量。

Page1





培训目标

学完本课程后，您应该能：

- 理解分类与标记的原理。
- 掌握分类与标记的方法。

Page2
HUAWEI

流量分类和标记

- 流量分类及标记是部署QoS 的基础
- 可以根据ACL、以及报文自身信息对流量进行分类
- 可以基于DSCP、IP Precedence、802.1P、MPLS EXP等信息对报文进行标记



 Page3 

流量分类是QoS技术的基础，是体现“差分”服务的基石。流量分类，就是将流量划分为多优先级或多个服务类，如使用IP报文头的ToS（Type of service，服务类型）字段的前三位（即IP优先级）来标记报文，可以将报文最多分成23 = 8类；若使用DSCP（Differentiated Services Codepoint，区分服务编码点，ToS域的的前6位），则最多可分成26 = 64类。在报文分类后，就可以将其它的QoS特性应用到不同的分类，实现基于类的拥塞管理、流量整形等。

对于流量的分类，几乎可以依据报文的任何信息，比如可以根据源IP地址、目的IP地址、源端口号、目的端口号、协议ID等进行流量的分类。

虽然流量分类几乎可以根据报文的任何信息进行，但是流量的标记则一般只对IP报文的ToS域进行标记。流量的标记主要的目的就是让其他处理此报文的应用系统或设备知道该报文的类别，并根据这种类别对报文进行一些事先约定了的处理（PHB）

例如，在网络的边界做如下分类和标记：

所有VoIP数据报文聚合为EF业务类，将报文的IP优先级标记为5，

或者将DSCP值标记为EF；

所有VoIP控制报文聚合为AF业务类，将报文的IP优先级标记为4，或者将DSCP值标记为AF31。

当报文在网络边界被标记分类之后，在网络的中间节点，就可以根据标记，对不同类别的流量给予差别服务了。例如对上述例子中的EF类业务保证时延和减少抖动，同时进行流量监管；对AF业务类在网络拥塞时仍然保证一定的带宽，等等。

流量分类

流量分类是按照一定的规则识别符合某类特征的报文，特征不同的报文享受到的服务不同。按照分类规则参考信息的不同，流量分类可以分为简单流分类和复杂流分类。

- 简单流分类是指采用简单的规则，如IP 报文头中的DSCP/IP-PRE值，MPLS报文的EXP域值，Vlan报文头中的802.1P 值对报文进行粗略的分类，以识别出具有不同优先级或服务等级特征的流量。
- 复杂流分类是指采用复杂的规则，如综合链路层、网络层、传输层信息（例如源MAC 地址、目的MAC 地址、源IP 地址、目的IP 地址、用户组号、协议类型或应用程序的TCP/UDP 端口号等）对报文进行精细的分类。通常在Diff-Serv 域的边界路由器上对流量进行复杂流分类。

 Page5

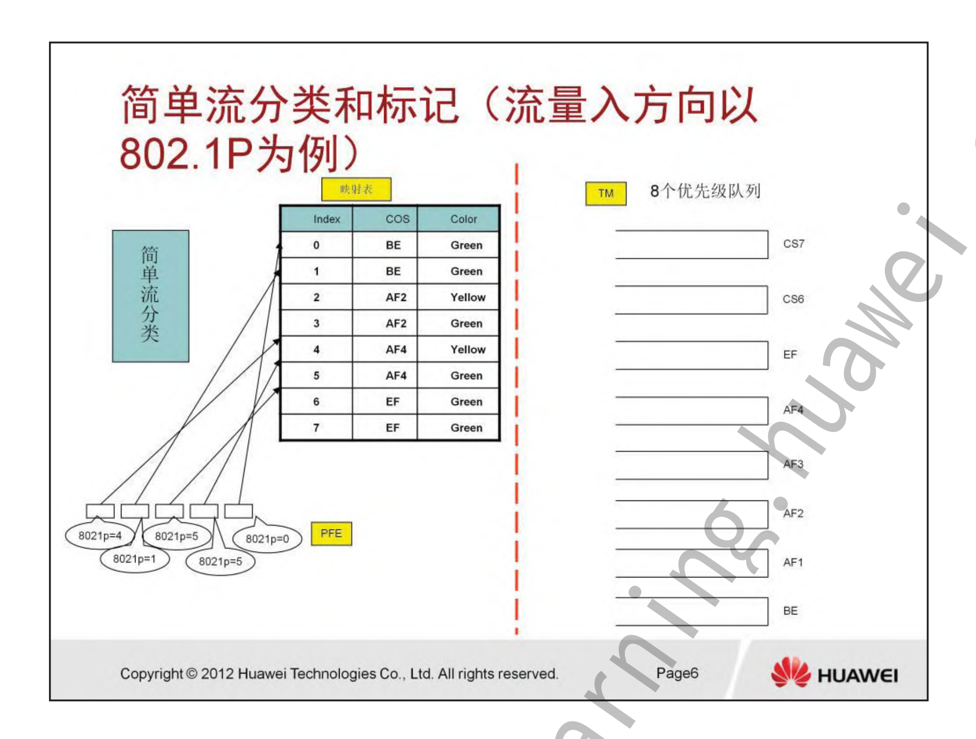


在采用Diff-Serv 模型实施QoS 时，需要路由器识别各种流，因此需要对报文进行流分类。

进行流分类是为了有区别地提供服务，它必须与某种流控或资源分配动作关联起来才有意义。具体采取何种流控动作，与所处的阶段以及网络当前的负载状况有关。例如，当报文进入网络时依据承诺速率对它进行监管；流出结点之前进行整形；拥塞时的队列调度管理；拥塞加剧时要采取拥塞避免措施等。

流量分类的规则可以按照IP报文头的信息（如 IP源地址、IP目的地址、协议类型、DSCP/IP-PRE、TCP/UDP 的端口号、TCP同步标志、报文分片的标志等）、二层报文信息（如： 源MAC地址、目的MAC地址、 三层报文的封装类型、 VLAN PRI等）、MPLS报文头信息（如EXP、 LSP等）来分类。





简单流量分类是将数据报文划分为多个优先级或多个服务类，如使用IP报文头的ToS（Type of service，服务类型）字段的前三位（即IP优先级）来标记报文，可以将报文最多分成8类；若使用DSCP（Differentiated Services Code Point，区分服务编码点，ToS域的前6位），则最多可分成64类。在报文分类后，就可以将其它的QoS特性应用到不同的分类，实现基于类的拥塞管理、流量整形等。

网络管理者可以设置报文简单流分类的策略，这个策略除可以包括IP报文的IP优先级或DSCP值、MPLS报文的EXP域值、802.1p的CoS值等带内信令。

通常于网络边界处对报文进行分类时，同时标记IP优先级或DSCP，这样，在网络的内部就可以简单的使用IP优先级或DSCP作为分类的标准。而队列技术如WFQ，CBWFQ就可以使用这个优先级来对报文进行不同的处理。下游（downstream）网络可以选择接收上游（upstream）网络的分类结果，也可以按照自己的分类标准对数据流量重新进行分类。

例如：在网络的边界做如下分类和标记：

所有VoIP数据报文聚合为EF业务类，将报文的IP优先级标记为5，或者将DSCP值标记为EF；

所有VoIP控制报文聚合AF业务类，将报文的IP优先级标记为4，或者将DSCP值标记为AF31。

当报文在网络边界被标记分类之后，在网络的中间节点，就可以根据标记，对不同类别的流量给予差别服务了。例如对上述例子中的EF业务类保证时延和减少抖动，同时进行流量监管；对AF业务类在网络拥塞时仍然保证一定的带宽，等等。

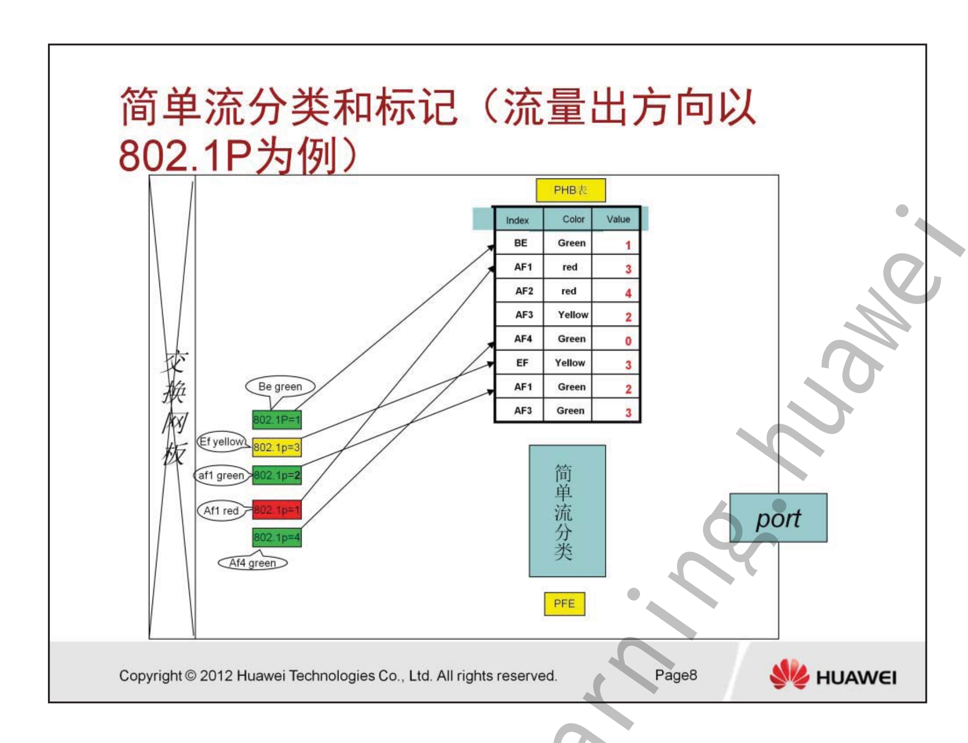
对于MPLS QoS，所谓标记就是MPLS报文中的EXP域进行设置。EXP域包括3位，虽然RFC 3032把它叫做实验（experimental）域，但它通常作为MPLS报文的CoS域，与IP网络的ToS域等效，用来区分数据流量的服务等级，以支持MPLS网络的DiffServ。

在IP网络，由IP报文的IP优先级或DSCP标识服务等级。但是对于MPLS网络，由于报文的IP头对LSR设备是不可见的，所以需要在MPLS网络的边缘对MPLS报文的EXP域进行标记。

缺省的情况下，在MPLS网络的边缘，将IP报文的IP优先级直接拷贝到MPLS报文的EXP域，但是在下面的情况下，如ISP不信任用户网络，或者ISP定义的差别服务类别不同于用户网络，则可以根据一定的简单分类策略，依据内部的服务等级重新设置MPLS报文的EXP域，而在MPLS网络转发的过程中保持IP报文的ToS域不变。

在MPLS网络的中间节点，根据MPLS报文的EXP域对报文进行分类，并实现拥塞管理，流量监管或者流量整形等PHB。





举例说明：图中的PHB表即华为路由器系统中的简单流分类映射表：

配置方法和察看方法见下一节：

```
<RT>display diffserv domain default
Diffserv domain name:default
 ……
 8021p-outbound be green map 1
 8021p-outbound af1 green map 2
 8021p-outbound af1 red map 3
 8021p-outbound af2 red map 4
 8021p-outbound af3 green map 3
 8021p-outbound af3 yellow map 2
 ……
```

简单流分类与标记在产品中的实现

华为路由器产品支持配置8个DS域（定义见注释）。

上行简单流分类，根据IP DSCP、MPLS EXP或802.1P将报文分为八种业务类型（CS7、CS6、EF、AF4—AF1、BE）、三种颜色（green、yellow、red），从而区分不同的业务（如，语音、视频、数据等）。在拥塞管理、队列调度时，不同业务进入不同的队列，得到差异化的调度。例如语音可以进入高优先级的PQ队列，保证低延时。上行若不做简单流分类，报文业务类型都为BE。

下行简单流分类，根据内部业务类型（CS7、CS6、EF、AF4—AF1、BE）、三种颜色（green、yellow、red），重新设置报文的IP DSCP、MPLS EXP或802.1P，实现了重标记的功能，重新标记IP DSCP、MPLS EXP或802.1P。下行未配置简单流分类时，IP DSCP、MPLS EXP或802.1P不做改变。

 Page9



域：定义了一组分类规则，规定了带不同优先级（dscp、mpls exp、802.1p等）的报文与内部优先级的映射关系，以及内部优先级与报文本身优先级的映射关系。

1、配置举例：

(1)在接口g3/0/0和g4/0/9,实现DSCP与EXP的映射。

配置diffserv域，名称为d1

[RT-0]diffserv domain d1

配置报文入方向的映射关系

[RT-0-dsdomain-d1]ip-dscp-inbound 34 phb af1 green

配置报文出方向的映射关系

[RT-0-dsdomain-d2]mpls-exp-outbound af1 green map 3

在接口下应用简单流分类

[RT-0-GigabitEthernet3/0/0]trust upstream d1

[RT-1-GigabitEthernet4/0/9]trust upstream d1

这样当DSCP为34的IP报文进入接口g3/0/0 后，根据简单流分类转换为路由器内部优先级af1（业务类型）， green颜色参与队列调度，流量管理等处理。当报文出g4/0/9时，简单流分类将根据其内部的优先级Af1和颜色green标记报文的exp值为3。

2、应用场景

IP，MPLS，VLAN报文跨域转换时，可以使用简单流分类实现DSCP/IP-PRE/EXP/802.1P之间的映射，并保证报文的服务等级不受变化。

3、华为路由器系统默认的简单流分类映射表:

```
<RT>display diffserv domain default
Diffserv domain name:default
 8021p-inbound 0 phb be green
 8021p-inbound 1 phb af1 green
 8021p-inbound 2 phb af2 green
 8021p-inbound 3 phb af3 green
 8021p-inbound 4 phb af4 green
 8021p-inbound 5 phb ef green
 8021p-inbound 6 phb cs6 green
 8021p-inbound 7 phb cs7 green
 8021p-outbound be green map 0
 8021p-outbound af1 green map 1
 8021p-outbound af1 yellow map 1
 8021p-outbound af1 red map 1
 8021p-outbound af2 green map 2
 8021p-outbound af2 yellow map 2
 8021p-outbound af2 red map 2
 8021p-outbound af3 green map 3
 8021p-outbound af3 yellow map 3
```

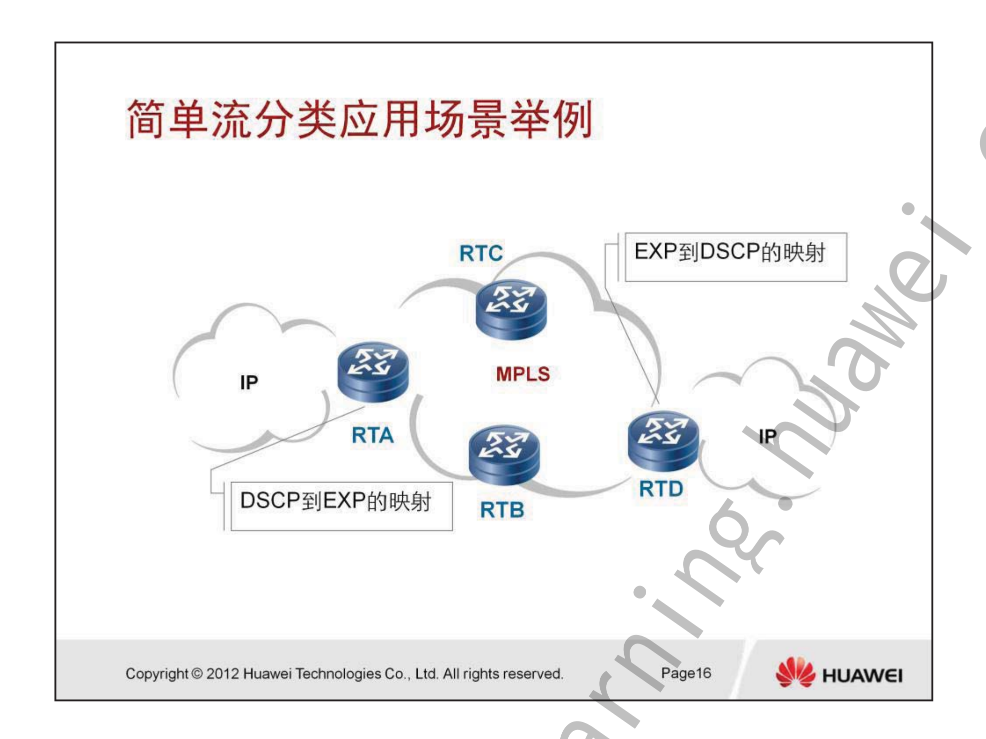
```
8021p-outbound af3 red map 3
 8021p-outbound af4 green map 4
 8021p-outbound af4 yellow map 4
 8021p-outbound af4 red map 4
 8021p-outbound ef green map 5
 8021p-outbound cs6 green map 6
 8021p-outbound cs7 green map 7
 ip-dscp-inbound 0 phb be green
 ip-dscp-inbound 1 phb be green
 ip-dscp-inbound 2 phb be green
 ip-dscp-inbound 3 phb be green
 ip-dscp-inbound 4 phb be green
 ip-dscp-inbound 5 phb be green
 ip-dscp-inbound 6 phb be green
 ip-dscp-inbound 7 phb be green
 ip-dscp-inbound 8 phb af1 green
 ip-dscp-inbound 9 phb be green
 ip-dscp-inbound 10 phb ef green
 ip-dscp-inbound 11 phb be green
 ip-dscp-inbound 12 phb af1 yellow
 ip-dscp-inbound 13 phb be green
 ip-dscp-inbound 14 phb af1 red
 ip-dscp-inbound 15 phb be green
 ip-dscp-inbound 16 phb af2 green
 ip-dscp-inbound 17 phb be green
 ip-dscp-inbound 18 phb af2 green
 ip-dscp-inbound 19 phb be green
 ip-dscp-inbound 20 phb af2 yellow
 ip-dscp-inbound 21 phb be green
```

```
ip-dscp-inbound 22 phb af2 red
ip-dscp-inbound 23 phb be green
ip-dscp-inbound 24 phb af3 green
ip-dscp-inbound 25 phb be green
ip-dscp-inbound 26 phb af3 green
ip-dscp-inbound 27 phb be green
ip-dscp-inbound 28 phb af3 yellow
ip-dscp-inbound 29 phb be green
ip-dscp-inbound 30 phb af3 red
ip-dscp-inbound 31 phb be green
ip-dscp-inbound 32 phb af4 green
ip-dscp-inbound 33 phb be green
ip-dscp-inbound 34 phb af4 green
ip-dscp-inbound 35 phb be green
ip-dscp-inbound 36 phb af4 yellow
ip-dscp-inbound 37 phb be green
ip-dscp-inbound 38 phb af4 red
ip-dscp-inbound 39 phb be green
ip-dscp-inbound 40 phb ef green
ip-dscp-inbound 41 phb be green
ip-dscp-inbound 42 phb be green
ip-dscp-inbound 43 phb be green
ip-dscp-inbound 44 phb be green
ip-dscp-inbound 45 phb be green
ip-dscp-inbound 46 phb ef green
ip-dscp-inbound 47 phb be green
ip-dscp-inbound 48 phb cs6 green
ip-dscp-inbound 49 phb be green
ip-dscp-inbound 50 phb be green
```

```
ip-dscp-inbound 51 phb be green
ip-dscp-inbound 52 phb be green
ip-dscp-inbound 53 phb be green
ip-dscp-inbound 54 phb be green
ip-dscp-inbound 55 phb be green
ip-dscp-inbound 56 phb cs7 green
ip-dscp-inbound 57 phb be green
ip-dscp-inbound 58 phb be green
ip-dscp-inbound 59 phb be green
ip-dscp-inbound 60 phb be green
ip-dscp-inbound 61 phb be green
ip-dscp-inbound 62 phb be green
ip-dscp-inbound 63 phb be green
ip-dscp-outbound be green map 0
ip-dscp-outbound af1 green map 10
ip-dscp-outbound af1 yellow map 12
ip-dscp-outbound af1 red map 14
ip-dscp-outbound af2 green map 18
ip-dscp-outbound af2 yellow map 20
ip-dscp-outbound af2 red map 22
ip-dscp-outbound af3 green map 26
ip-dscp-outbound af3 yellow map 28
ip-dscp-outbound af3 red map 30
ip-dscp-outbound af4 green map 34
ip-dscp-outbound af4 yellow map 36
ip-dscp-outbound af4 red map 38
ip-dscp-outbound ef green map 46
ip-dscp-outbound cs6 green map 48
ip-dscp-outbound cs7 green map 56
```

```
mpls-exp-inbound 0 phb be green
 mpls-exp-inbound 1 phb af1 green
 mpls-exp-inbound 2 phb af2 green
 mpls-exp-inbound 3 phb af3 green
 mpls-exp-inbound 4 phb af4 green
 mpls-exp-inbound 5 phb ef green
 mpls-exp-inbound 6 phb cs6 green
 mpls-exp-inbound 7 phb cs7 green
 mpls-exp-outbound be green map 0
 mpls-exp-outbound af1 green map 1
 mpls-exp-outbound af1 yellow map 1
 mpls-exp-outbound af1 red map 1
 mpls-exp-outbound af2 green map 2
 mpls-exp-outbound af2 yellow map 2
 mpls-exp-outbound af2 red map 2
 mpls-exp-outbound af3 green map 3
 mpls-exp-outbound af3 yellow map 3
 mpls-exp-outbound af3 red map 3
 mpls-exp-outbound af4 green map 4
 mpls-exp-outbound af4 yellow map 4
 mpls-exp-outbound af4 red map 4
 mpls-exp-outbound ef green map 5
 mpls-exp-outbound cs6 green map 6
 mpls-exp-outbound cs7 green map 7
 atm-inbound ubr 0 phb be green
 atm-inbound ubr 1 phb be green
 atm-inbound cbr 0 phb ef green
 atm-inbound cbr 1 phb ef green
 atm-inbound rt-vbr 0 phb af4 green
```

```
atm-inbound rt-vbr 1 phb af4 yellow
atm-inbound nrt-vbr 0 phb af2 green
atm-inbound nrt-vbr 1 phb af2 yellow
atm-inbound abr 0 phb af1 green
atm-inbound abr 1 phb af1 yellow
atm-inbound oam-cell phb ef green
atm-outbound be green map 1
atm-outbound af1 green map 1
atm-outbound af1 yellow map 1
atm-outbound af1 red map 1
atm-outbound af2 green map 0
atm-outbound af2 yellow map 1
atm-outbound af2 red map 1
atm-outbound af3 green map 1
atm-outbound af3 yellow map 1
atm-outbound af3 red map 1
atm-outbound af4 green map 0
atm-outbound af4 yellow map 1
atm-outbound af4 red map 1
atm-outbound ef green map 0
atm-outbound cs6 green map 0
atm-outbound cs7 green map 0
ppp-inbound control phb ef green
<RT>
```

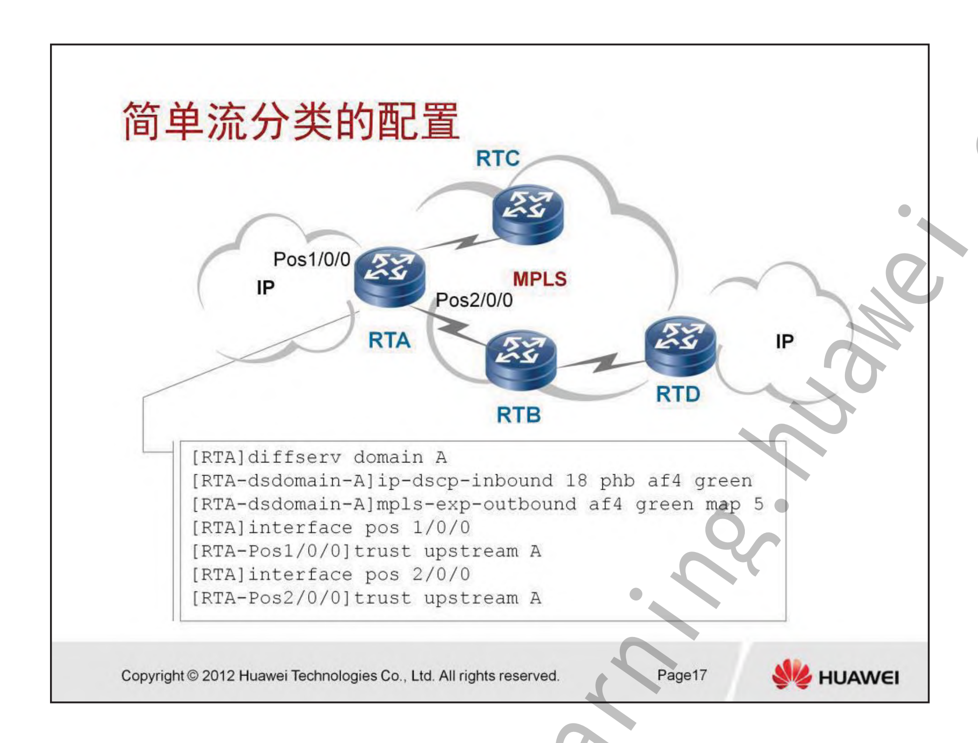


简单流分类是指根据IP报文的IP优先级或DSCP值、MPLS报文的EXP域值、VLAN报文的802.1p值，将报文划分为多个优先级或多个服务等级。配置基于简单流分类的流量策略可以将一种网络流量中的优先级映射到另外一种网络流量中，使流量在另外一种网络中按照原来的优先级传送。

简单流分类的应用场景：在IP，MPLS，VLAN报文跨域转换时，可以使用简单流分类实现DSCP/IP-PRE/EXP/802.1P之间的映射，并保证报文的服务等级不受变化。

简单流分类通常配置在网络的核心位置。

简单流分类不止应用在物理端口，还可以用于逻辑端口。在企业组网中逻辑端口有更加广阔的应用，



在本配置实例中，RTA、RTB、RTC、RTD运行MPLS协议，RTA与RTD分别连接有IP网络。假定MPLS网络中四台路由器上的MPLS配置已经完成，IP流量从RTA到RTD能进行MPLS转发，MPLS流量出RTD时，能转换成IP流，并且保持QoS不变。

RTA的基本配置流程如下：

1、首先建立Differserv域；

2、配置IP报文DSCP 18对应的PHB为AF4并将报文标记为绿色；

3、配置MPLS的服务等级为AF4的绿色报文对应的MPLS EXP为5；

4、分别在RTA的IP和MPLS接口使能Differserv 域。

Diffserv domain ds-domain-name，定义DS域并进入DS域视图。

ip-dscp-inbound dscp-value phb service-class [color] 命令用来配置当前域的上行IP报文的DSCP值对应的服务等级，并将报文着色，缺省情况下为green。

dscp-value：指定上行IP报文的DSCP值。整数形式，取值范围是0～63。

Phb service-class：指定对应的服务等级，取值为EF、AF1、

AF2、AF3、AF4、BE、CS6或CS7。

color：指定报文标记的颜色，取值为green、yellow或red。

mpls-exp-outbound service-class color map exp命令用来配置当前域下行报文的服务等级和颜色标记对应的EXP域值。

service-class：指定服务等级，取值为EF、AF1、AF2、AF3、AF4、BE、CS6和CS7。

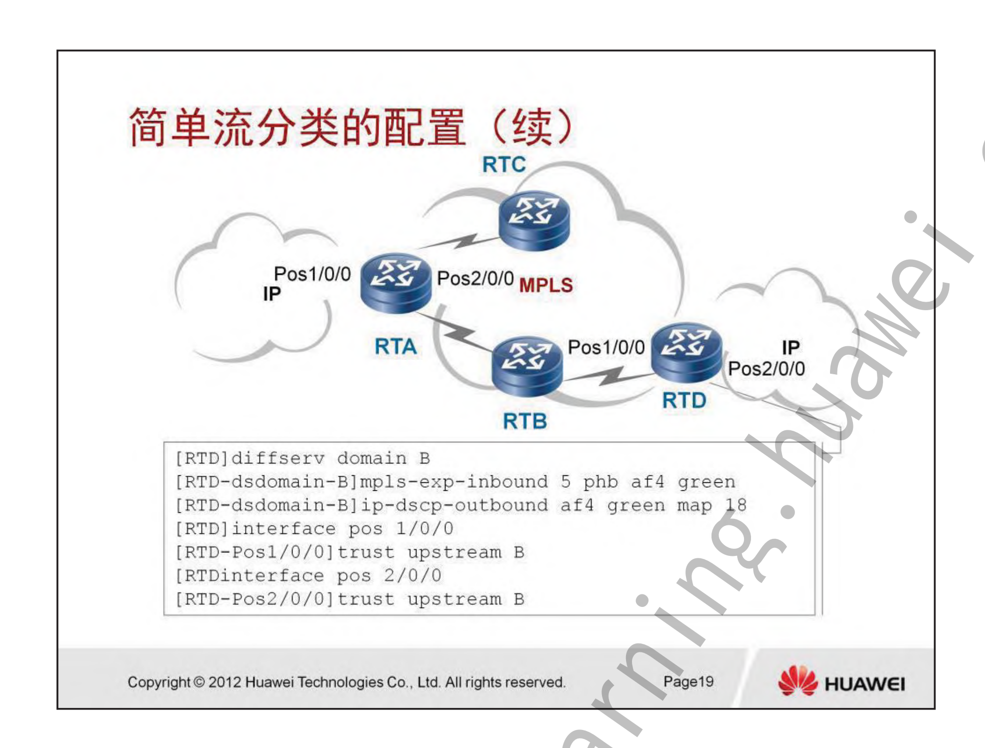
color：指定报文标记的颜色，取值为green、yellow或red。

map exp：指定对应的MPLS报文的EXP域值。整数形式，取值范围是0～7。

trust upstream { ds-domain-name | default } 命令用来在接口上绑定DS域，使其支持简单流分类，以便对IP报文、MPLS报文的优先级进行修改。

ds-domain-name：DS区域名称，字符串形式，长度范围是1～8。DS区域(DS domain)由一组采用相同的服务提供策略和实现了相同PHB组集合的相连DS节点组成。

default：指定默认DS域。



RTA的IP流进入MPLS网络后按MPLS标签转发，由于RTD连接有MPLS和IP两个网络，所以在RTD上需要配置MPLS QoS到IP QoS的映射。具体配置步骤如下：

1、首先建立Differserv域；

2、配置MPLS报文EXP 5对应的PHB为AF4并将报文标记为绿色；

3、配置IP的服务等级为AF4的绿色报文对应的DSCP为18；

4、分别在RTD的IP和MPLS接口使能Differserv 域。

mpls-exp-inbound exp phb service-class [color]命令用来配置当前域上行MPLS报文的EXP域值的对应的服务等级，并将报文着色，缺省情况下为green。

exp：指定MPLS报文的EXP域值。整数形式，取值范围是0～7。

phb service-class：指定对应的服务等级，取值为EF、AF1、AF2、AF3、AF4、BE、CS6或CS7。

color：指定报文标记的颜色，取值为green、yellow或red。服务等级为CS6、CS7、EF和BE时，只能将报文着色为green。

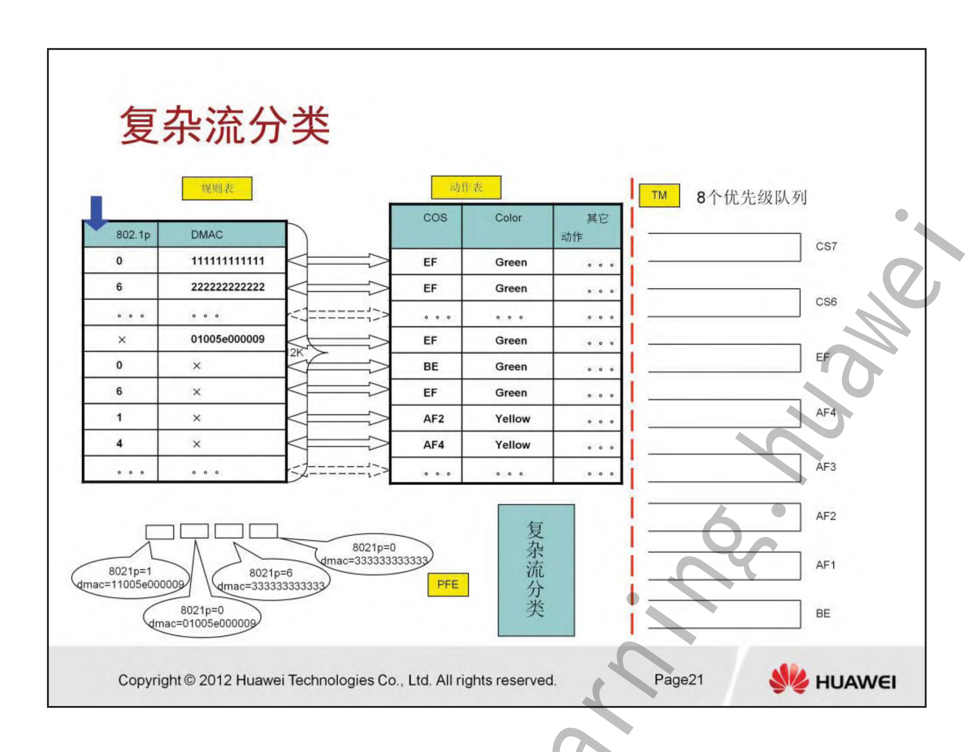
ip-dscp-outbound service-class color map dscp-value命令用来配

置当前域的下行IP报文服务等级和颜色对应的DSCP值。

service-class：指定下行IP报文的服务等级，取值为EF、AF1、AF2、AF3、AF4、BE、CS6或CS7。

color：指定报文标记的颜色，取值为green、yellow或red。

map dscp-value：指定对应的DSCP值，整数形式，取值范围是0～63。



复杂流分类是指根据五元组（源地址、源端口号、协议号码、目的地址、目的端口号）等报文信息对报文进行分类（一般的分类依据都局限在封装报文的头部信息，使用报文内容作为分类的标准比较少见），缺省应用于网络的边缘位置。报文进入边缘节点时，网络管理者可以灵活配置分类规则。分类的结果是没有范围限制的，它可以是一个由五元组（源地址、源端口号、协议号码、目的地址、目的端口号）确定的狭小范围，也可以是匹配某网段的所有报文。

复杂流分类通过提取报文信息，如报文优先级、源IP、目的IP、源MAC、目的MAC、802.1p、报文封装类型等等，组成关键字去匹配规则表，然后通过匹配规则得到一个索引，再根据索引查动作表，将报文映射为内部优先级，除了映射内部优先级外，复杂流分类还可以支持流量监管(CAR)、PBR（ Policy-based Routing ）、重标记，报文过滤、采样、镜像等其它动作。

复杂流分类在产品中的实现

在实现复杂流分类时分为两个部分：规则部分和动作部分。

当处理报文时，根据报文中用来分类的字段信息组成关键字，查找规则表；如果报文能匹配上规则部分，则根据查找结果确定该规则对应的动作表，确定该报文应该执行何种动作。如果报文没有匹配上任何一条规则，则报文不做分类按普通报文正常转发。

ACL（Access Control List）：访问控制列表。用于复杂流分类的规则部分。

 Page22 

产品在实现复杂流分类时分为两个部分：复杂流分类的规则部分与复杂流分类的动作部分。

配置举例：

1、创建复杂流分类的规则

定义一个名为c1的类。

<Quidway> system view

[Quidway] traffic classifier c1

[Quidway-classifier-c1]

配置一条匹配规则 dscp = 1。

[Quidway-classifier-c1] if-match dscp 1

2、创建复杂流分类的动作部分

定义一个名为b1的流行为。

<Quidway> system view

[Quidway] traffic behavior b1

[Quidway-behavior-b1]

配置一个重标记的动作。

[Quidway-behavior-b1] remark dscp ef

3、将规则部分与动作部分结合起来，组成复杂流分类的流策略

定义一个名为p1的策略。

<Quidway> system view

[Quidway] traffic policy p1

在流策略p1中配置符合流分类c1的报文采用流行为b1。

<Quidway> system view

[Quidway] traffic policy p1

[Quidway-trafficpolicy-p1] classifier c1 behavior b1 。

4、将该策略应用到接口上，复杂流分类功能生效

将流量策略p1应用到接口GigabitEthernet 1/0/0的出方向上。

<Quidway> system view

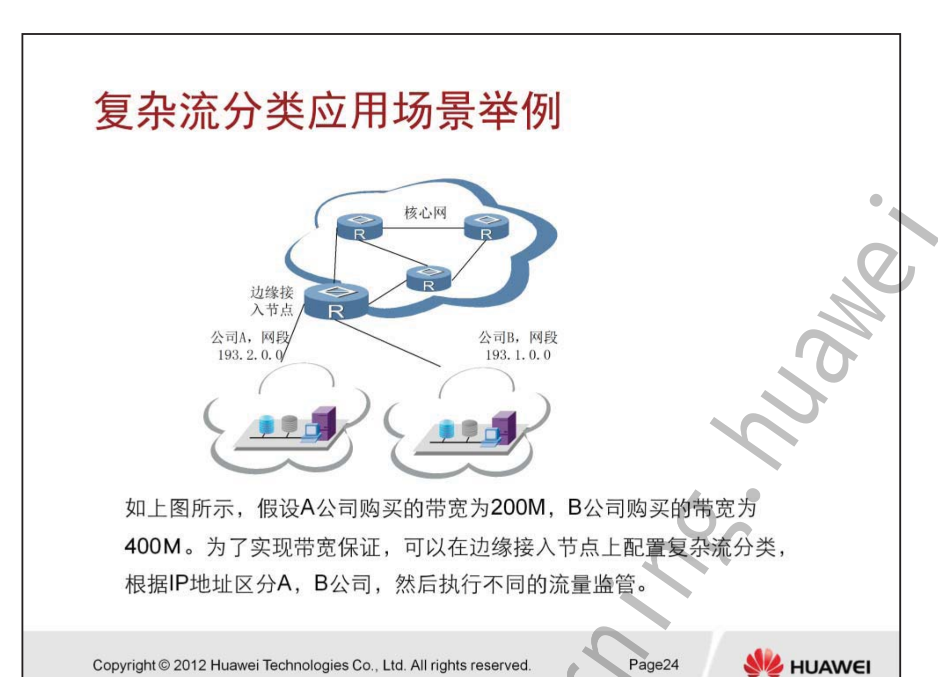
[Quidway] interface gigabitethernet1/0/0

[Quidway-GigabitEthernet1/0/0] traffic-policy p1 outbound

复杂流分类在处理报文时，根据报文中用来分类的字段信息组成关键字，查找规则表，根据查找结果确定动作表的索引，再根据动作表索引查找具体的动作，确定报文应该执行何种动作。这样当不同的规则关联上不同的动作时就可以提供差分服务，实现网络中对于不同报文做不同处理的需求。

对于复杂流分类的规则部分，目前的产品实现中可以支持根据以太报文头中的源MAC地址、目的MAC地址、报文链路层承载的协议号、带TAG报文的优先级进行分类；支持根据IPv4报文的IP优先级/DSCP/ToS域值、源IP地址前缀、目的IP地址前缀、IP报文承载的协议号、分片标志，TCP SYN标志、TCP/UDP源端口号或端口范围、TCP/UDP目的端口号或端口范围进行分类；支持根据IPv6报文的IP优先级/DSCP/ToS域值、源IP地址前缀、目的IP地址前缀、IP报文承载的协议号、TCP/UDP源端口号或端口范围、TCP/UDP目的端口号或端口范围进行分类。

对于复杂流分类的动作部分，目前的产品实现中可以支持报文过滤、流量监管（CAR）、报文重标记、PBR、报文镜像、报文采样、报文防攻击检查（uRPF）。



注释：复杂流分类应用场景举例的相关配置见流量监管的配置举例。



问 题

什么是流量分类?

流量分类包括哪些分类方法?

简单流分类是依据报文的什么信息来分类?

复杂流分类是依据报文的什么信息来分类?

 Page25 HUAWEI



答案:

- 流量分类是按照一定的规则识别符合某类特征的报文，特征不同的报文享受到的服务不同。
- 按照分类规则参考信息的不同，流量分类可以分为简单流分类和复杂流分类;
- 简单流分类是根据报文的优先级，如IP 报文头中的DSCP/IP-PRE信息，MPLS报文头中的MPLS EXP信息，二层报文头中的802.1P 信息来分类的。
- 复杂流分类是根据更多的报文信息，如报文优先级、源IP、目的IP、源MAC、目的MAC、协议类型等信息来分类的。



谢谢

www.huawei.com


流量监管与整形
www.huawei.com
Copyright © 2012 Huawei Technologies Co., Ltd. All rights reserved.
HUAWEI



前 言

流量整形是对报文的速率进行控制，使报文以均匀的速率发送出去。流量监管为了使有限的网络资源可以更好地为用户服务，可以对特定用户的业务流进行监管，使其适应分配给它的那部分网络资源。

 Page1





培训目标

学完本课程后，您应该能：

- 理解流量监管与整形的原理。
- 掌握流量监管与整形的方法。

 Page2 HUAWEI

流量监管介绍

流量监管（Traffic-policing）是一种在入接口或出接口应用的对进入路由器的某流量进行限制的流量管理技术。

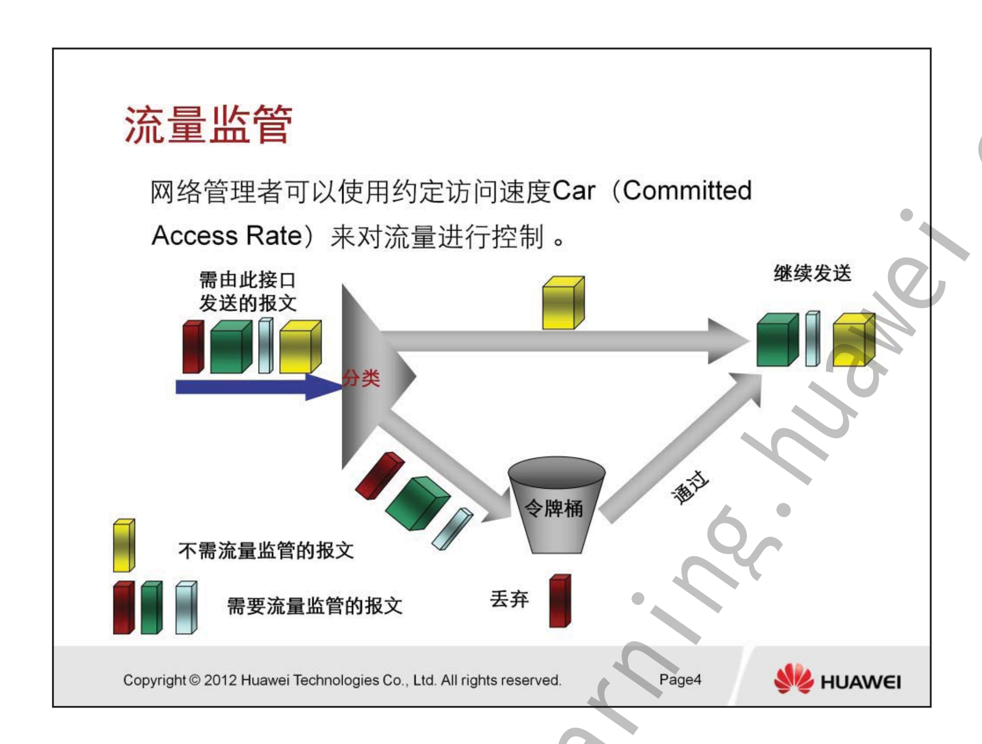
对于 ISP 来说，对用户送入网络中的流量进行控制是十分必要的。对于企业网，对某些应用的流量进行控制也是一个有力的控制网络状况的工具。

流量监管的典型应用是监督进入网络的某一流量的规格，把它限制在一个合理的范围之内，或者对超出的部分流量进行“惩罚”，以保护网络资源和运营商的利益。在报文满足一定的条件时，如某个连接的报文流量过大，流量监管就可以对该报文采取不同的处理动作，例如丢弃报文。

 Page3



从高速链路向低速链路传输数据时，带宽会在低速链路接口处出现瓶颈，导致数据丢失严重，特别是会影响到低时延要求的数据如语音等。流量监管（traffic policing）的典型作用是限制进入或流出某一网络的某一连接的流量与突发。在报文满足一定的条件时，如某个连接的报文流量过大，流量监管就可以对该报文采取不同的处理动作，例如丢弃报文，或重新设置报文的优先级等。通常的用法是使用CAR来限制某类报文的流量，例如限制HTTP报文不能占用超过50%的网络带宽。



CAR 利用 TB （Tocken Bucket – 令牌桶 ）进行流量控制，上图是CAR的处理过程，首先报文被分类，如果报文需要流量监管，则进入令牌桶中进行处理，如果令牌桶中有足够的令牌可以用来发送报文，则报文可以通过并被继续发送下去。如果令牌桶中的令牌不满足报文的发送条件，则报文被丢弃。这样就可以对某类报文的流量进行控制。

CAR可以对特定流量进行流量监管，对超出限额的流量进行丢弃或者重新标记。

首先，根据预先设置的匹配规则来对报文进行分类，如果是无须进行流量监管的报文，就直接发送，不需要经过令牌桶的处理。

其次，如果是需要进行流量控制的报文，则会进入令牌桶中进行处理。我们定义包长度为B，令牌数量为TB：

如果进入令牌桶处理的包长度B –TB<0，报文能拿到令牌则报文是绿色的（不管对此颜色的报文采取什么处理行为如丢弃、转发等），同时令牌桶中减少相应报文长度的令牌，TB=TB-B，反之拿不到令牌的报文是红色的，令牌不减少（不论对红色报文执行转发还是其他的处理行为）。需要说明的是如果把红色报文的处

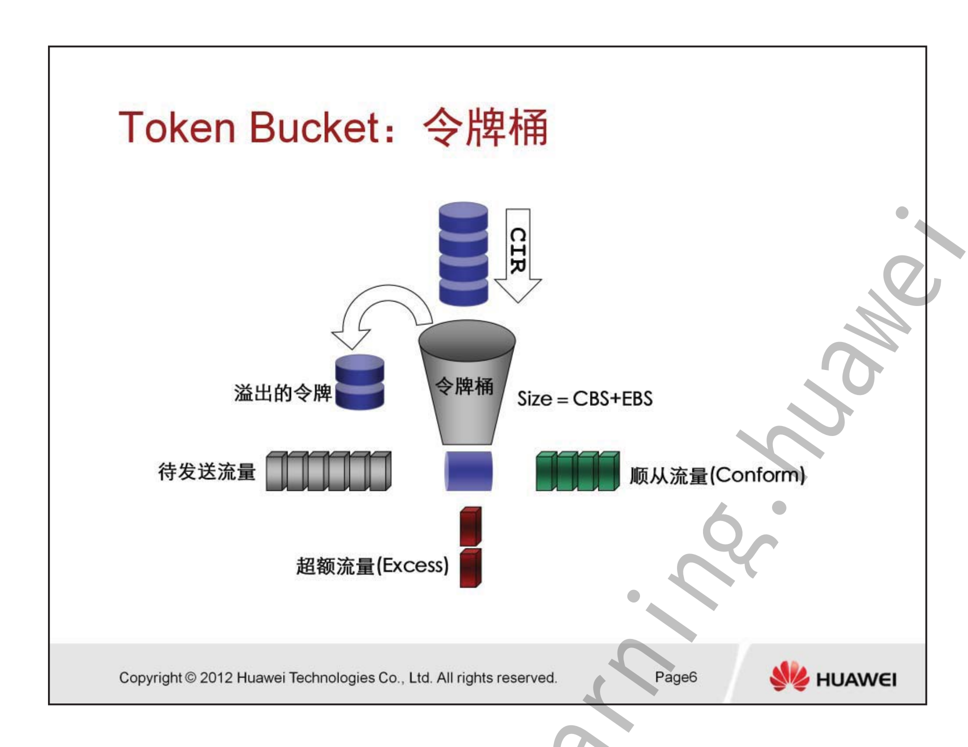
理行为配置成PASS，虽然拿不到令牌，但是报文还是可以送出。例如，如果进入令牌桶处理的报文长度B为800bits,TB= 30000bits,这时30000-800>0,所以报文是绿色的，同时令牌数量TB=30000-800；否则报文是红色的，TB不减少。

当令牌桶中没有令牌的时候，报文将不能被发送，只有等到桶中生成了新的令牌，报文才可以发送，这就可以限制报文的流量只能是小于等于令牌生成的速度，达到限制流量的目的。令牌桶按用户设定的速度向桶中放置令牌，并且用户可以设置令牌桶的容量。

在实际应用中，VRP的CAR不仅可以用来进行流量控制，还可以进行报文的标记（mark）或重新标记（re-mark）。具体来讲就是CAR可以设置IP报文的优先级或修改IP报文的优先级，达到标记报文的目的。例如，当报文符合流量特性的时候，可以设置报文的优先级为5，当报文不符合流量特性的时候，可以丢弃，也可以设置报文的优先级为1并继续进行发送。这样，后续的处理可以尽量保证不丢弃优先级为5的报文，在网络不拥塞的情况下，也发送优先级为1的报文，当网络拥塞时，首先丢弃优先级为1的报文，然后才丢弃优先级为5 的报文。

CAR可以为不同类别的报文设置不同的流量特性和标记特性。即，首先对报文进行分类，从而使不同类别的报文有不同的流量特性和标记特性。

此外，CAR的策略还可以进行串联处理。例如，可以对所有的报文限制一个总的流量，然后在总的流量中，再限制部分报文的流量符合某个流量特性。CAR能在出接口和入接口上生效。



令牌桶用来评估流量速率是否超过了规定值，以采取相应的措施。

令牌桶中装的是令牌而不是分组，每隔Δ t时间产生一个令牌，放入令牌桶中，令牌桶满后，新产生的令牌将被丢弃。

一个令牌表示可发送一个字节或一定数量字节的数据报文；当报文到来的时候，如果令牌桶中有足够的令牌用来发送数据，则报文通过，同时令牌的数量按令牌长度作相应的减少；如果令牌不足以发送一个数据报文的话，则这个报文被丢弃，令牌数量不变。

以令牌桶中令牌的数量是否满足报文的转发作为依据，评估的结果有两种：顺从（Conform）或超标（Excess）。

评估流量时，令牌桶的参数设置包括：

1、平均速率(Committed Information Rate)：向桶中放置令牌的速率

2、突发尺寸(Committed Burst Size)：令牌桶的容量，每次突发所允许的最大流量尺寸，设置的突发尺寸必须大于最大报文长度

为了测量更复杂的情况，实施更灵活的调控策略，可以设置两个令牌桶。例如流量策略TP（Traffic Policing）中有三个参数：承诺信息速率CIR（Committed Information Rate）；承诺突发尺寸



CBS（Committed Burst Size）；超出突发尺寸EBS（Excess Burst Size）。它使用了两个令牌桶，每个桶投放令牌的速率一样，均为CIR，只是尺寸不同——分别为CBS和EBS，简称C桶和E桶，代表所允许的不同突发级别。每次测量时，依据“C桶有足够的令牌”、“C桶令牌不足，但E桶足够”以及“C桶和E桶都没有足够的令牌”的情况，可以分别实施不同的流控策略。

流量监管的具体实现

单桶单速率流量监管：一个令牌桶，容量是CBS，一个填充令牌的速率CIR。当有 B 字节的报文传过来的时候，根据桶的当前容量来对这个报文进行处理。

双桶单速率流量监控：两个令牌桶，一个的容量是CBS，一个的容量是EBS，一个填充令牌的速率CIR,两个令牌桶使用同一个填充速率。当有B 字节的报文传过来的时候，根据两个桶的当前容量来对这个报文进行处理。

双桶双速率流量监控：两个令牌桶，一个的容量是CBS，一个的容量是PBS。这两个令牌桶分别使用两个填充令牌的速率，一个填充速率是CIR，一个填充速率是 PIR。当有 B 字节的报文传过来得时候，根据两个桶的当前容量来对这个报文进行处理。

　Page8



1、单令牌桶主要由两个参数构成：

– CIR（Committed Information Rate）：承诺信息速率，表示向令牌桶中投放令牌的速率，即长时间的端口平均速率

– CBS（Committed Burst Size）：承诺突发尺寸，用来决定在部分流量超过CIR之前的最大突发流量，即为令牌桶的容量（深度）。设置的突发尺寸必须大于报文的最大长度

2、单速双桶，主要由三个参数构成

– 承诺信息速率CIR（Committed Information Rate）：表示向C 桶中放置令牌的速率，即C 桶允许的流的平均速度

– 承诺突发尺寸CBS（Committed Burst Size）：表示C 桶的容量，即每次突发C桶所允许的最大的流量尺寸

– 额外突发尺寸EBS（Extra Burst Size）：表示E 桶的容量，即每次突发E 桶所允许的最大的流量尺寸

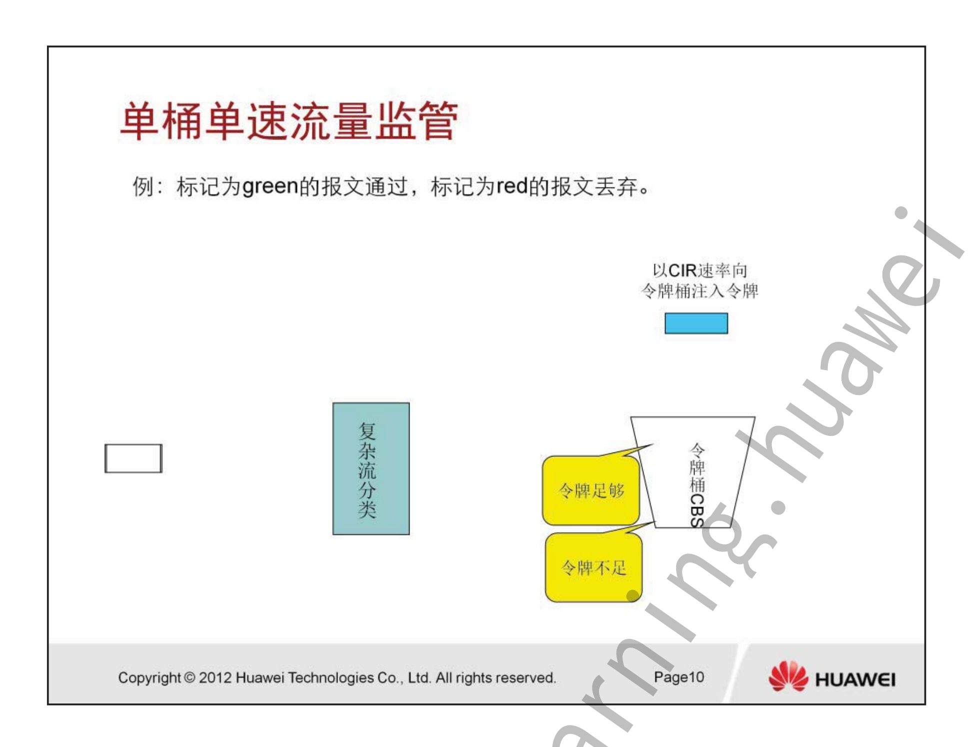
3、双速双桶，主要由四个参数构成

– 承诺信息速率CIR（Committed Information Rate）：表示向C 桶中放置令牌的速率，即C 桶允许的流的平均速度

– 承诺突发尺寸CBS（Committed Burst Size）：表示C 桶的容量，即每次突发C桶所允许的最大的流量尺寸

– 峰值信息速率PIR（Peak Information Rate）：表示向P 桶中放置令牌的速率，即P 桶允许的流的平均速度

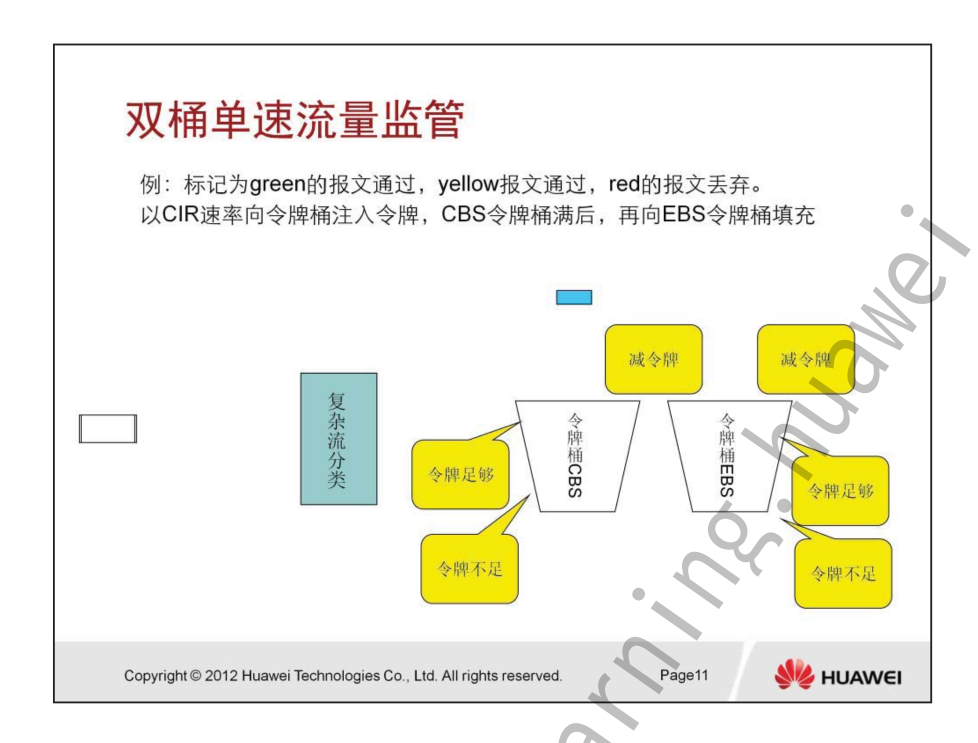
– 过度突发尺寸PBS（Peak Burst Size）：表示P 桶的容量，即每次突发P 桶所允许的最大的流量尺寸



流量监管单桶单速配置举例:

在接口GE6/0/0的入方向上配置流量监管。设置报文正常流量为1M，对于green报文允许通过，red报文丢弃。

```
<Quidway> system view
[Quidway] interface gigabitethernet6/0/0
[Quidway-GigabitEthernet6/0/0] qos car cir 1000 cbs 10000 pbs 0 green pass  red discard inbound
```



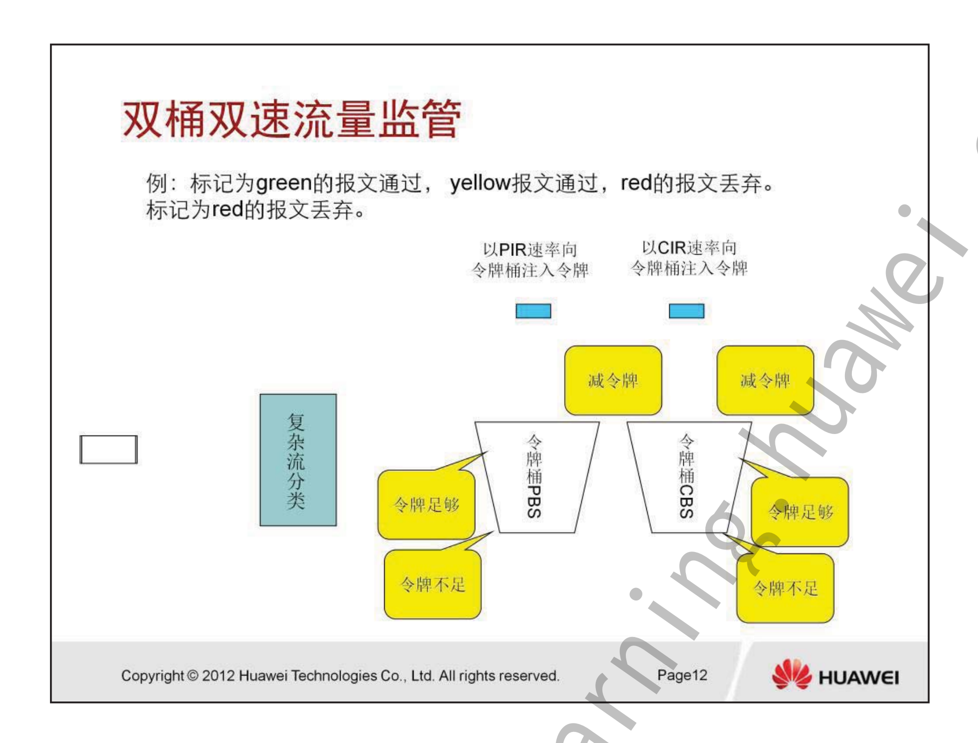
流量监管双桶单速配置举例：

在接口GE6/0/0的入方向上配置流量监管。设置报文正常流量为1M。

<Quidway> system view

[Quidway] interface gigabitethernet6/0/0

[Quidway-GigabitEthernet6/0/0] qos car cir 1000 cbs 10000 pbs 10000 green pass yellow pass red discard inbound



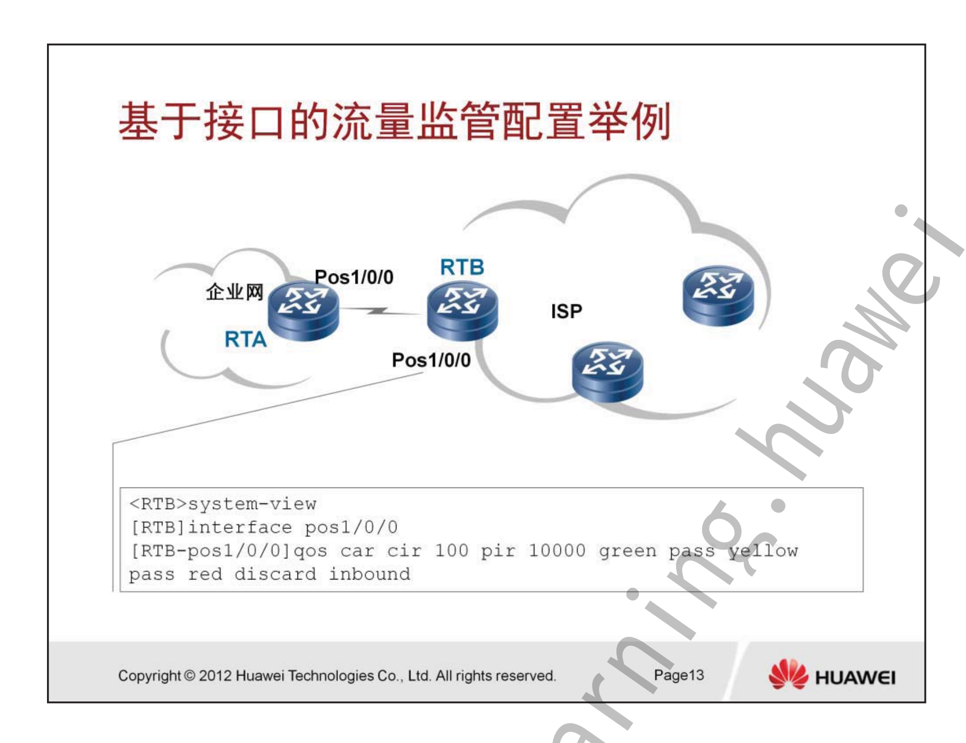
流量监管双桶双速配置举例:

在接口GE6/0/0的入方向上配置流量监管。设置报文正常流量为1M，突发流量为2M。

<Quidway> system view

[Quidway] interface gigabitethernet6/0/0

[Quidway-GigabitEthernet6/0/0] qos car cir 1000 pir 2000 cbs 10000 pbs 20000 green pass yellow pass red discard inbound



基于接口的流量监管是指对进入该接口的所有流量进行控制，而不区分具体报文的类型，一般应用于网络核心路由器。

本例中在RTB的POS1/0/0端口上配置CAR来实现流量监管，对从POS1/0/0接口的所有流量进行控制，CIR为100，PIR为10000，绿色和黄色正常通过，红色报文丢弃。具体的配置解释如下。

1、执行命令system-view，进入系统视图。

2、执行命令interface interface-type interface-number，进入接口视图。

3、执行命令qos car cir cir-value [pir pir-value] [cbs cbs-value pbs pbs-value] [green { discard | pass } [yellow { discard | pass } [red { discard | pass }]]] { inbound | outbound }，在接口上配置CAR。

cir cir-value：指定承诺信息速率(Committed Information Rate)，即保证能够通过的速率。整数形式，取值范围是100～10000000，单位是kbit/s。

pir pir-value：指定峰值速率PIR(Peak Information Rate)，即最大能够通过的速率。整数形式，取值范围是100～10000000，单位

是kbit/s。参数pir-value的值不应小于已经配置的cir-value的值。

cbs cbs-value：指定承诺突发尺寸(Committed Burst Size)，即瞬间能够通过的承诺流量，即第一个令牌桶的深度(假定该桶为A桶)。整数形式，取值范围是64～33554432，单位是byte。缺省值与配置的cir-value有关。若cir-value<=10000kbit/s，则cbs-value的缺省值为10000byte；若cir-value>10000kbit/s，cbs-value的缺省值等于cir-value的值，单位是byte。

pbs pbs-value：指定过度突发尺寸(Peak Burst Size)，即瞬间能够通过的峰值流量，即第二个令牌桶的深度(假定该桶为B桶)。整数形式，取值范围是0～33554432，单位是byte。缺省值与pir-value有关。如果不设置pir-value，则pbs-value缺省值为0；如果设置了pir-value，并且pir-value<10000kbit/s，则pbs-value缺省值为10000byte；如果设置了pir-value，并且pir-value>=10000kbit/s，则pbs-value缺省值等于pir-value，单位是byte。

pass、discard：指定对着色为某种颜色(green、yellow、red)的报文采取的动作，分别是通过、丢弃。

green：数据包的流量符合承诺信息速率时对数据包采取的动作，缺省值为pass。

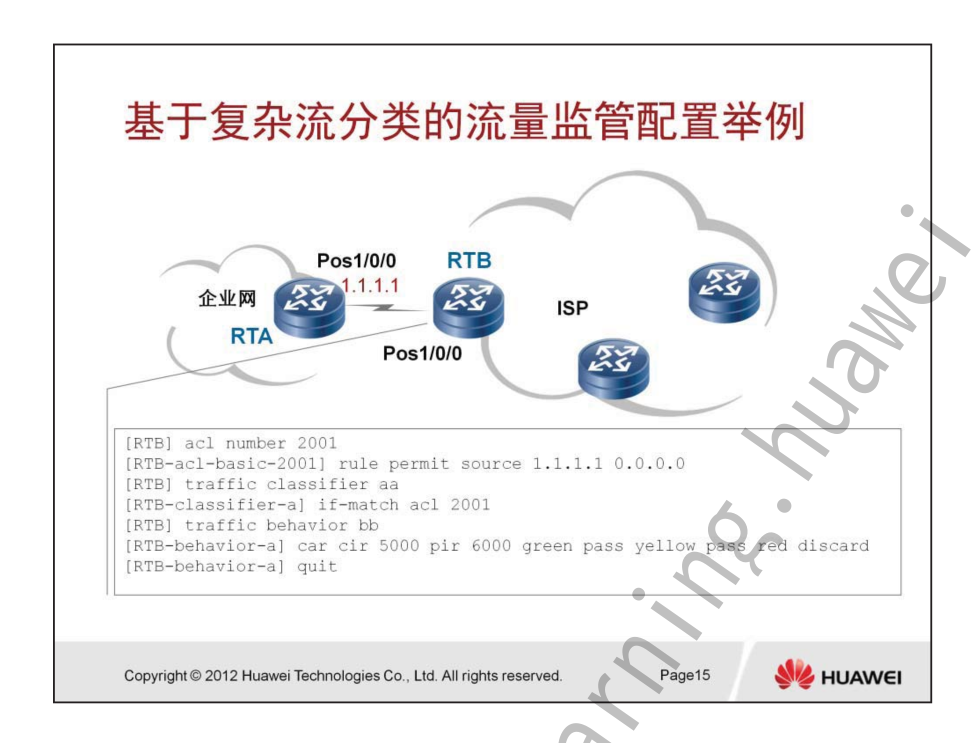
yellow：数据包的流量超过承诺信息速率但小于峰值速率时对数据包采取的动作，缺省值为pass。

red：数据包的流量超过峰值速率时对数据包采取的动作，缺省值为discard。

inbound：在报文的入(上行)方向配置流量监管。

outbound：在报文的出(下行)方向配置流量监管。

此例中CAR部分配置了CIR，PIR两个参数，而默认情况下不配置CBS和PBS参数也可以得到缺省的CBS和PBS，在trTCM算法中提到，trTCM算法是通过CIR，PIR，CBS，PBS四个参数来定义的，所以此例采用的算法为trTCM双速率三色标记算法，且CAR工作在色盲模式。



复杂流分类是指根据五元组(源地址、源端口号、协议号码、目的地址、目的端口号)等报文信息对报文进行分类，缺省应用于网络的边缘位置。

如果需要对进入接口的满足特定条件的某一类或几类报文进行流量控制，而非所有报文。可以考虑将复杂流分类与流量控制行为相结合，配置基于复杂流分类的流量监管策略，然后应用于该接口。

本例中在RTB上定义对从源为1.1.1.1网段发来的报文做流量监管，CIR为5000，PIR为6000，绿色和黄色报文正常通过，红色报文丢弃。

基于复杂流分类的流量监管配置解释如下：

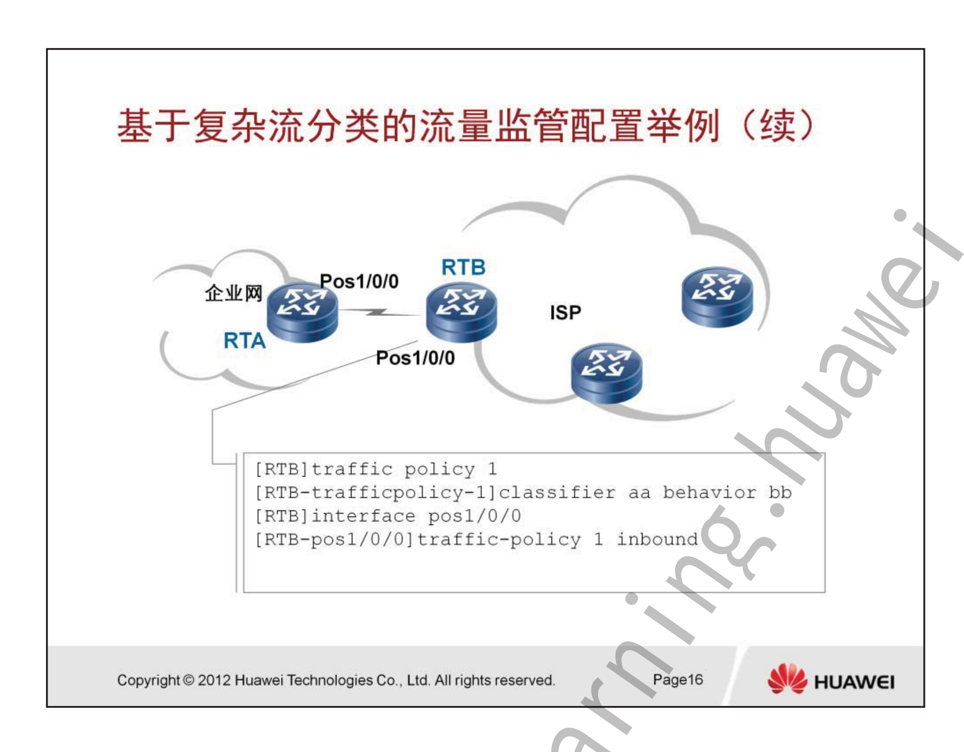
1、执行命令system-view，进入系统视图。

2、执行命令traffic classifier classifier-name [operator { and | or }]，定义流分类并进入类视图。

3、请根据实际情况对路由器的匹配规则进行定义。本例定义报文发送者的IP地址。

4、执行命令traffic behavior behavior–name，定义行为进入流行为视图。

5、执行命令car cir cir-value [pir pir-value] [cbs cbs-value pbs pbs-value] [green { discard | pass } [yellow { discard | pass } [red { discard | pass }]]]，配置流量监管动作。



6、执行命令traffic policy policy-name，定义流量策略并进入策略视图。

7、执行命令classifier classifier-name behavior behavior-name，在流量策略中为类指定采用的行为。

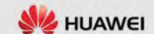
8、执行命令interface interface-type interface-number，进入接口视图。

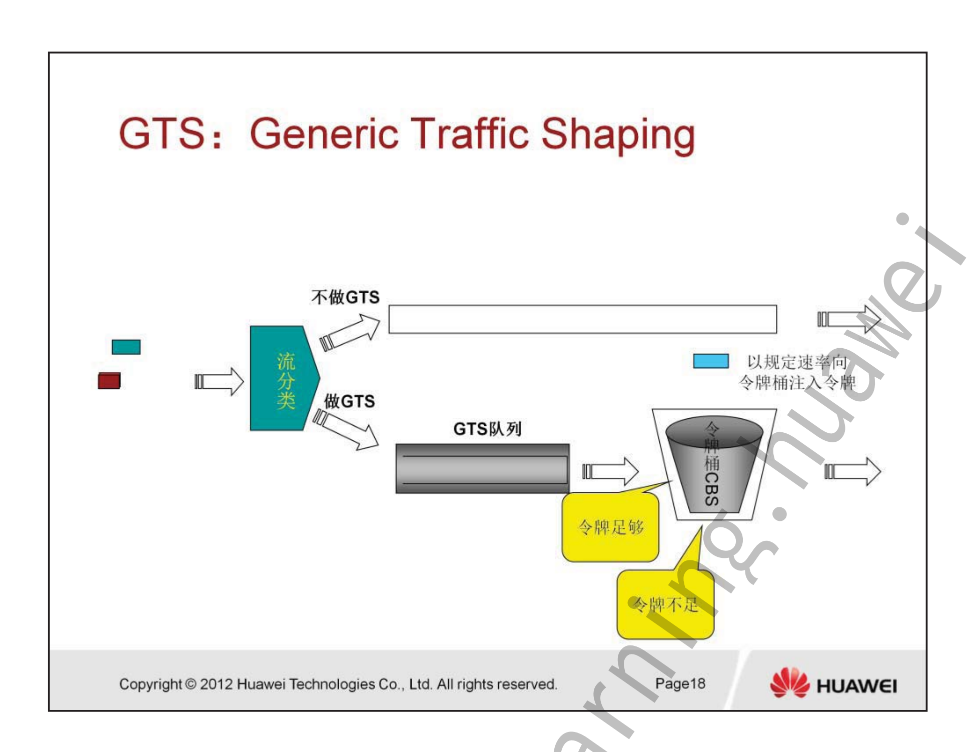
9、traffic-policy policy-name { inbound | outbound } [link-layer]，在接口应用流量策略。

流量整形介绍

流量整形（traffic shaping）的典型作用是限制流出某一网络的某一连接的流量与突发，使这类报文以比较均匀的速度向外发送。流量整形通常使用缓冲区或队列和令牌桶来完成，当报文的发送速度过快时，首先在缓冲区或队列进行缓存，在令牌桶的控制下，再均匀地发送这些被缓冲的报文。

流量整形通常采用的技术有：Generic Traffic Shaping（通用流量整形，简称GTS），Line Rate（物理接口总速率限制，简称LR）。它们可以对不规则或不符合预定流量特性的流量进行整形，以利于网络上下游之间的带宽匹配。





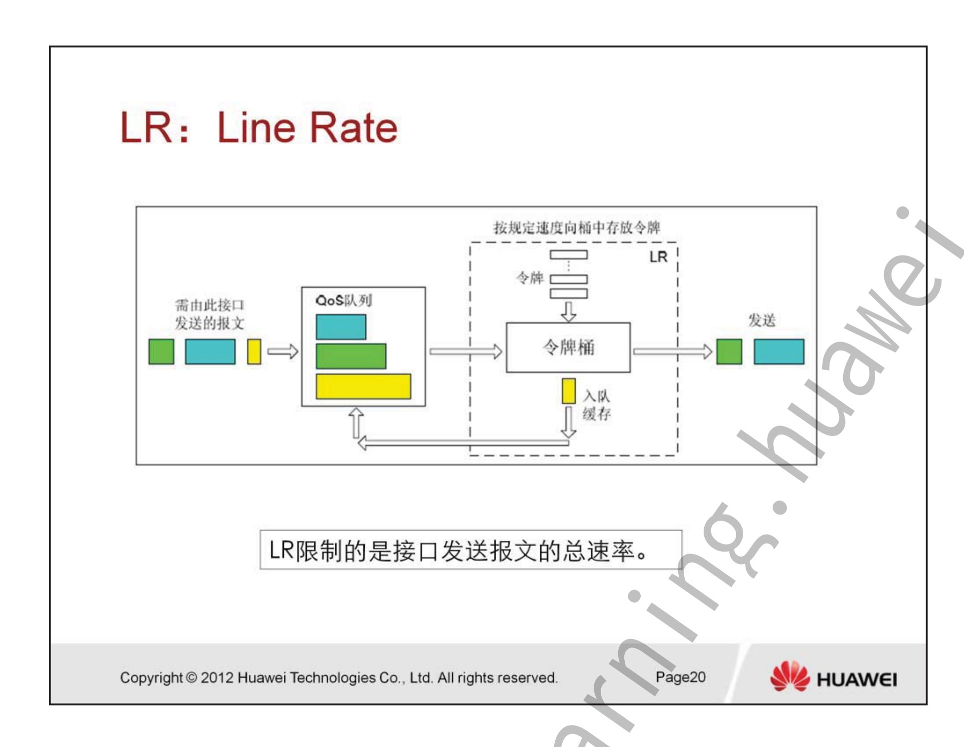
通用流量整形（简称GTS）可以对不规则或不符合预定流量特性的流量进行整形，以保证网络上下游之间的带宽匹配，避免拥塞发生。

GTS与CAR一样，都采用了令牌桶技术来控制流量。GTS与CAR的主要区别在于：利用CAR进行报文流量控制时，对不符合流量特性的报文进行丢弃；而GTS对于不符合流量特性的报文则是进行缓冲，减少了报文的丢弃，同时满足报文的流量特性。

GTS的基本处理过程如上图所示，其中用于缓存报文的队列称为GTS队列。

GTS可以对接口上指定的报文流或所有报文进行整形。当报文到来的时候，首先对报文进行分类，如果报文不需要进行GTS处理，就继续发送，不需要经过令牌桶的处理；流量整形的令牌桶的构成同CAR一样，如果报文需要进行GTS处理，则与令牌桶中的令牌进行比较，进入令牌桶处理的包长度B –TB<0则报文被发送，否则报文被缓存，等到令牌桶中有足够的令牌时继续发送报文。令牌桶按用户设定的速度向桶中放置令牌，如果令牌桶中有足够的令牌可以用来发送报文，则报文直接被继续发送出去，同时，

令牌桶中的令牌量按报文的长度做相应的减少。当令牌桶中的令牌少到报文不能再发送时，报文将被缓存入GTS队列中（队列是FIFO队列），此队列与接口上的FIFO不是同一个队列，当然队列有一定的长度（以包为单位），当需要缓存的报文个数大于队列长度时报文因无法缓存而丢弃。当GTS队列中有报文的时候，GTS按一定的周期从队列中取出报文进行发送，每次发送都会与令牌桶中的令牌数作比较，令牌数足够则发送，令牌数不够就继续缓存。另外，GTS也允许有突发。GTS只能在出接口上生效。



物理接口总速率限制（简称LR）可以在一个物理接口上，限制接口发送报文（包括紧急报文）的总速率。

LR的处理过程仍然是采用令牌桶进行流量控制。如果用户在路由器的某个接口上配置了LR，规定了流量特性，则所有经由该接口发送的报文首先要经过LR的令牌桶进行处理。如果令牌桶中有足够的令牌可以用来发送报文，则报文可以发送。如果令牌桶中的令牌不满足报文的发送条件，则报文进入QoS队列进行拥塞管理。这样，就可以对通过该物理接口的报文流量进行控制。LR的处理过程如图所示。

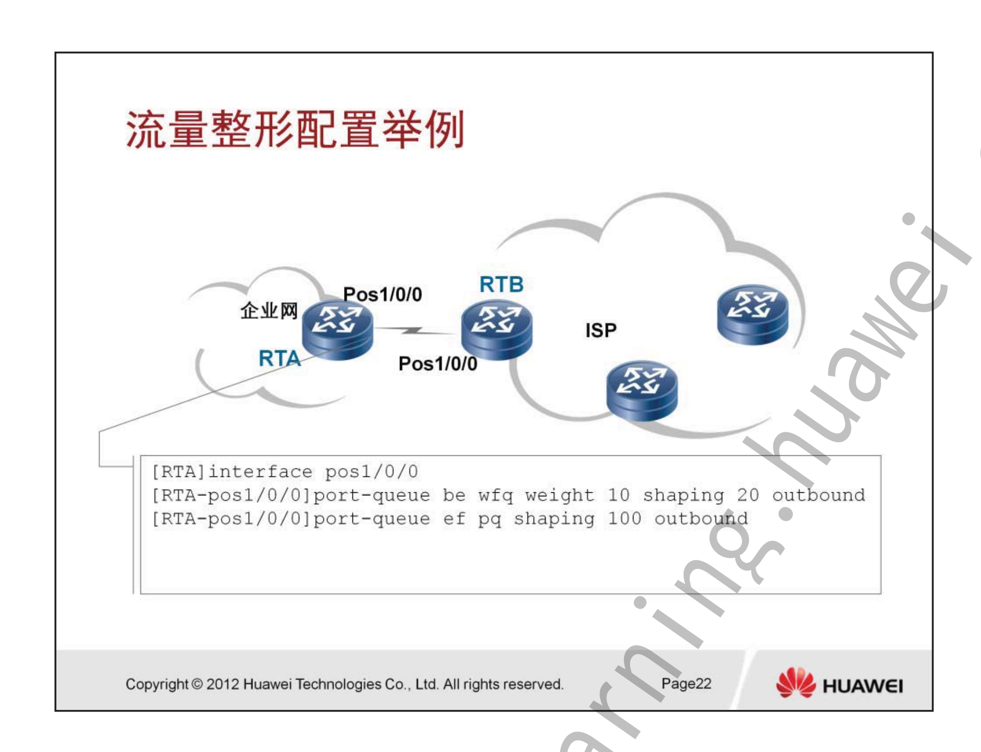
同样的，由于采用了令牌桶控制流量，当令牌桶中积存有令牌时，可以允许报文的突发性传输。当令牌桶中没有令牌的时候，报文将不能被发送，只有等到桶中生成了新的令牌，报文才可以发送，这就可以限制报文的流量只能是小于等于令牌生成的速度，具有限制流量，同时允许突发流量通过的目的。

LR能够限制在物理接口上通过的所有报文，CAR和GTS在IP层实现，对不经过IP层处理的报文不起作用。LR与GTS比较，LR不但能够对超过流量限制的报文进行缓存，而且还使报文进入了Qos

队列机制进行处理，所以队列调度机制更灵活。

在用户只要求对所有报文限速时，使用LR所需的配置操作简单。

对于网络建设投资者，可以对客户隐藏实际带宽，客户只能严格按所购买的带宽来使用。



在华为高端系列路由器中使用port-queue命令用来配置接口出(下行)方向的QoS服务等级承诺信息速率和峰值速率，以及队列调度的优先级。

port-queue cos-value { { pq | wfq weight weight-value | lpq} | shaping { shaping-value | shaping-percentage shaping-percentage-value } | port-wred wred-name } * outbound

参数说明

cos-value：指定配置的流队列优先级。取值可以是af1、af2、af3、af4、be、cs6、cs7、ef。

weight-value：流队列调度的权重。整数形式，取值范围是1～100。

shaping-value：整形速率，表示配置的接口带宽，等于峰值信息速率PIR的取值。整数形式，取值范围是0～1000，单位为Mbps。

shaping-percentage-value：整形速率百分比。表示每个流队列的整形速率占配置的端口输出带宽的百分比。整数形式，取值范围是0～100。

pq | wfq | lpq：配置该队列的调度方式。pq为绝对优先级队列调度；wfq为加权公平队列调度；lpq为低优先级调度。

三种队列调度的优先级次序为：

PQ队列的优先级高于WFQ队列的优先级。

WFQ队列的优先级高于LPQ队列的优先级。

高优先级的队列可以抢占低优先级队列的带宽。

wred-name：配置该队列的WRED模板。字符串形式，长度范围是1～31。

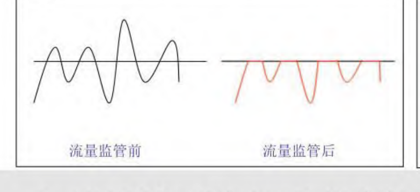
outbound：修改该接口8个CQ队列出（下行）方向的调度参数。

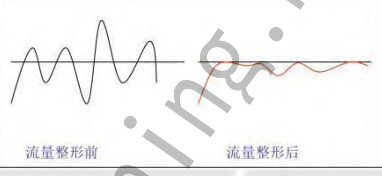
本例中在RTA路由器的pos1/0/0出接口配置流量整形，配置EF服务类采用PQ调度算法，PIR的峰值速率所占接口带宽的百分比为20%；BE服务类采用WFQ调度，配置权重为10，PIR的峰值速率所占接口带宽的百分比为10%。

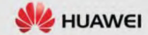
流量监管与流量整形的区别

GTS、LR与CAR三者均采用了令牌桶技术来控制流量。它们的主要区别在于：在进行报文流量控制时，CAR对超过流量限制的报文进行丢弃；而GTS则将报文缓存在GTS队列中。 相较于GTS，LR不但能够对超过流量限制的报文进行缓存，并且可以利用QoS丰富的队列来缓存报文。

类型	优点	缺点
流量整形	丢包量少，抗突发强	时延大，需要Buffer资源缓存报文
流量监管	不需要Buffer资源，时延小	丢包量大，抗突发弱





 Page24 



问 题

什么是流量监管?

流量监管的实现方法包括哪些?

流量整形的作用是什么?

流量监管与流量整形的区别是什么?

答案:

- 流量监管（Traffic-policing）是一种在入接口或出接口应用的对进入路由器的某流量进行限制的流量管理技术。
- 流量监管的具体实现包括单桶单速、双桶单速、双桶双速方法。
- 流量整形（traffic shaping）的典型作用是限制流出某一网络的某一连接的流量与突发，使这类报文以比较均匀的速度向外发送
- 流量整形与流量监管的主要区别：在进行报文流量控制时，流量监管是对超过流量限制的报文进行丢弃；而流量整形则将超过流量限制的报文缓存在队列中。

谢谢

www.huawei.com


拥塞管理与拥塞避免
www.huawei.com
Copyright © 2012 Huawei Technologies Co., Ltd. All rights reserved.
HUAWEI



前 言

当网络中间歇性的出现拥塞，时延敏感业务要求得到比非时延敏感业务更高质量的QoS服务时，需要进行拥塞管理。拥塞避免是指通过监视网络资源（如队列或内存缓冲区）的使用情况，在拥塞发生或有加剧的趋势时主动丢弃报文，通过调整网络的流量来解除网络过载的一种流量控制机制。

 Page1





培训目标

学完本课程后，您应该能：

- 理解拥塞管理与拥塞避免的原理。
- 掌握拥塞管理与拥塞避免的方法。

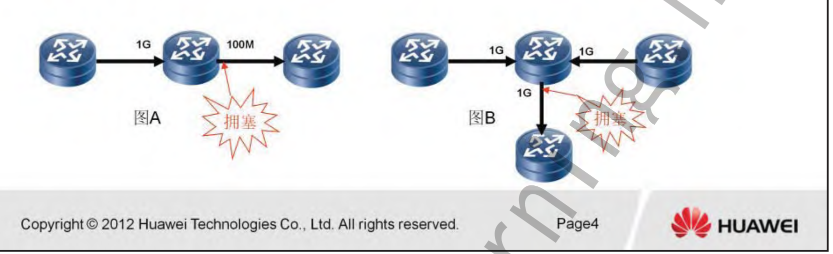
 Page2 HUAWEI


目 录
拥塞管理
拥塞避免
Copyright © 2012 Huawei Technologies Co., Ltd. All rights reserved.
Page3
HUAWEI

拥塞与拥塞管理

流量从高速端口流向低速端口会在低速端口上产生拥塞，如图A；
流量从多个端口流向同一个端口会在汇聚端口上产生拥塞，如图B

拥塞管理是指网络在发生拥塞时，如何进行管理和控制。处理的方法是使用队列调度技术。将所有要从一个接口发出的报文进入多个队列，按照各个队列的优先级进行处理。通过适当的队列调度机制，可以优先保证某种类型的报文的QoS 参数，例如带宽、时延、抖动等。



Page4　HUAWEI

传统网络所面临的服务质量问题，主要是由网络拥塞引起的。所谓拥塞，是指由于供给资源的相对不足而造成服务速率下降（引入了额外的延迟）的一种现象。

对于网络单元，当数据包到达的速度大于该接口发送数据包的速度时，在该接口处就会产生拥塞。如果没有足够的存储空间来保存这些分组，它们其中的一部分就会丢失。数据包的丢失又可能会导致发送该数据包的主机或路由器因超时而重传此数据包，这将导致恶性循环。

当拥塞发生时，多个报文会同时竞争使用资源， 导致得不到资源的某些业务报文被丢弃，尤其不能保证关键业务的带宽、时延、抖动等 Qos参数。此时如何制定一个资源的调度策略决定报文转发的处理次序，就是拥塞管理的中心内容。对于拥塞管理，一般采用排队技术，它包括队列的创建，决定报文的队列归属的流分类，以及队列间的调度策略。

最初的队列调度策略只有FIFO（先进先出），后来为了满足不同业务的需求，设计了多种调度策略。

队列调度机制由两部分组成：硬件队列和软件队列。硬件队列也

叫发送队列（Transmit Queue，TxQ），接口驱动在逐个传输数据包的时候需要使用这个队列，这个队列是FIFO队列。软件队列根据QoS的要求把数据包调度到硬件队列，软件队列可以使用多种调度方式。

数据包在硬件队列已满的情况下才进入软件队列调度。

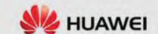
硬件队列的长度跟接口的带宽设置有关，接口带宽大，传输时延就小，因此队列长度可以设得长一些，反之，队列长度要设得小一些。合理设置硬件队列长度非常关键，硬件队列长度过长会影响软件队列执行策略的效果，因为硬件队列是使用FIFO机制进行调度的；硬件队列长度太短会影响调度的效率，导致链路利用率降低和CPU占用过高。

队列技术

当拥塞发生时，多个报文会同时竞争使用资源， 导致得不到资源的某些业务报文被丢弃，尤其不能保证关键业务的带宽、时延、抖动等 QoS参数。此时如何制定一个资源的调度策略决定报文转发的处理次序，就是拥塞管理的中心内容。对于拥塞管理，一般采用队列调度技术，常见的队列调度技术有以下几种：

- FIFO：First In First Out，先进先出队列
- RR：Round Robin，轮询队列
- WRR：Weight Round Robin，按权重轮询队列
- PQ：Priority Queuing，优先级队列
- CQ：Custom Queuing，自定义队列
- WFQ：Weighted Fair Queuing，加权公平队列

 Page6



常用的队列调度机制如下：

1、FIFO：First In First Out，先进先出队列

2、RR：Round Robin，轮询队列

3、WRR：Weight Round Robin，按权重轮询队列

4、PQ：Priority Queuing，优先级队列

5、CQ：Custom Queuing，自定义队列

6、WFQ：Weighted Fair Queuing，加权公平队列

队列是一个比较容易理解的概念，我们在日常生活中也用到类似技术。例如我们去电影院买票的时候，大家排成几队顺序买票，排在前面的先拿到票；有时突然冲出一个人跑到队伍的最前面拿出证件马上就拿到了票，这类人属于特权阶级需要优先处理，后面的人只能等这类人买完票才能继续排队买票。

队列调度机制是QoS中非常重要的一个技术，是拥塞管理机制。在出接口发生拥塞时，通过适当的队列调度机制，可以优先保证某种类型的报文的QoS参数，例如带宽、时延、抖动等。我们这里所说的队列是指出队列，其作用是在接口有能力发送报文之前

先将报文在内存中保留下来，直到接口可以继续发送报文，所以队列调度机制都是在出端口发生拥塞情况下产生作用，另外一个主要作用就是将报文重新排序（FIFO除外）。

主要的队列调度包括：FIFO、RR、WRR、PQ、CQ、WFQ，我们可以从分类、丢弃策略、单一队列内的调度方式、队列之间的调度方式、队列数目和队列长度各方面来比较学习这些队列技术。



FIFO是队列机制中最简单的。每个接口上只有一个FIFO队列，表面上看FIFO队列并没有提供什么QoS保证，实则不然。既然只有一个队列，自然不需要考虑把什么类型的报文放入到哪个队列的问题，也不用考虑下一个报文从哪个队列拿包、怎么拿、拿多少的问题，即FIFO无需流分类、调度机制，而且因为按顺序取报文，FIFO无需对报文重新排序。简化了这些程序其实也就提高了对时延的保证。

FIFO关心的就是队列长度问题，队列长度会影响到时延、抖动、丢包率。因为队列长度是有限的，有可能被填满，这就涉及到该机制的丢弃原则，FIFO使用Tail Drop机制。如果定义了较长的队列长度，那么队列不容易填满，被丢弃的报文也就少了，但是队列长度太长了会出现时延的问题，一般情况下时延的增加会导致抖动也增加；如果定义了较短的队列，时延的问题可以得到解决，但是发生Tail Drop的报文就变多了。类似的问题其它排队方法也存在。

Tail Drop机制简单的说就是如果该队列已经满了，那么后续进入的报文被丢弃，而没有什么机制来保证后续的报文可以挤掉已经

在队列内的报文。

优点：

1、简单。

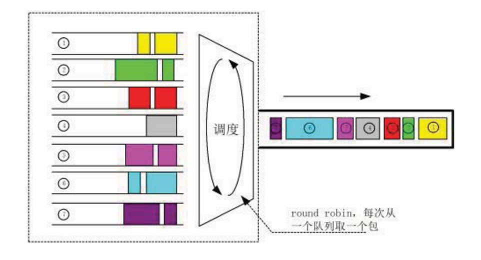
缺点：

1、没有公平性，不同的流之间不互相隔离，当某一个流的带宽太大的时候会占用其他流的带宽，并且造成其他流的时延增加。

2、当拥塞发生的时候，FIFO对一部分报文进行丢弃。TCP的连接发现有丢包后，会降低传输的速度，来主动的避免拥塞，但是UDP是非连接的，不降低发送速率。导致FIFO中TCP和UDP的报文不平衡，TCP的流量太低。

3、一条流的突发流量可能占用全部buffer，将其他的流量全都阻断。

RR：Round Robin





RR是Round Robin的缩写，是一种简单的调度方法，采用轮询的方式，对多个队列进行调度RR以环形的方式轮询多个队列。如果轮询的队列不为空，则从该队列取走一个报文；如果该队列为空，则直接跳过该队列，调度器并不等待。

 Page10 HUAWEI

优点：

1、隔离了不同的流，实现了队列之间对带宽的平等利用

2、剩余带宽能够被其他队列平均分配

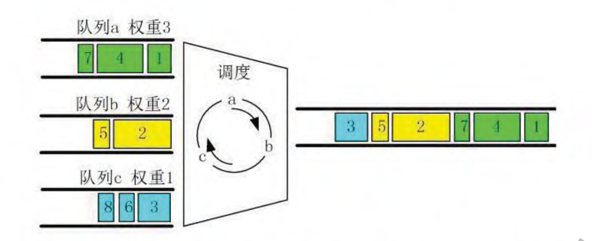
缺点：

1、无法设置队列占用带宽的权重；

2、当不同队列中的报文长度不一的时候，调度不准确；

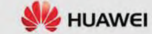
3、当调度速率低的时候，时延和抖动的问题比较突出，比如一个包到达一个空队列，而这个队列刚刚被调度完毕，则这个包要等到其他全部的队列调度完才能取得出接口的机会，这样会导致抖动比较大，但是如果调度速度非常高，则这种时延可以忽略，RR在高速路由器内部有很多应用。

WRR：Weight Round Robin



WRR（Weighted Round Robin）主要针对RR不能设置权重的不足，在轮询的时候，每个队列享受的机会和该队列的权重成比例。WRR最初是针对固定包长（ATM）设计的调度算法。WRR对于空的队列直接跳过，调度一周结束的时间变短，因此当某个队列的流量小的时候，剩余带宽能够被其他队列按照比例占用。

 Page11



优点：

1、能够按照权重来分配带宽，某个队列的剩余带宽能够为其他队列公平占用，低优先级的队列同样能够得到调度，不存在饥饿的问题。

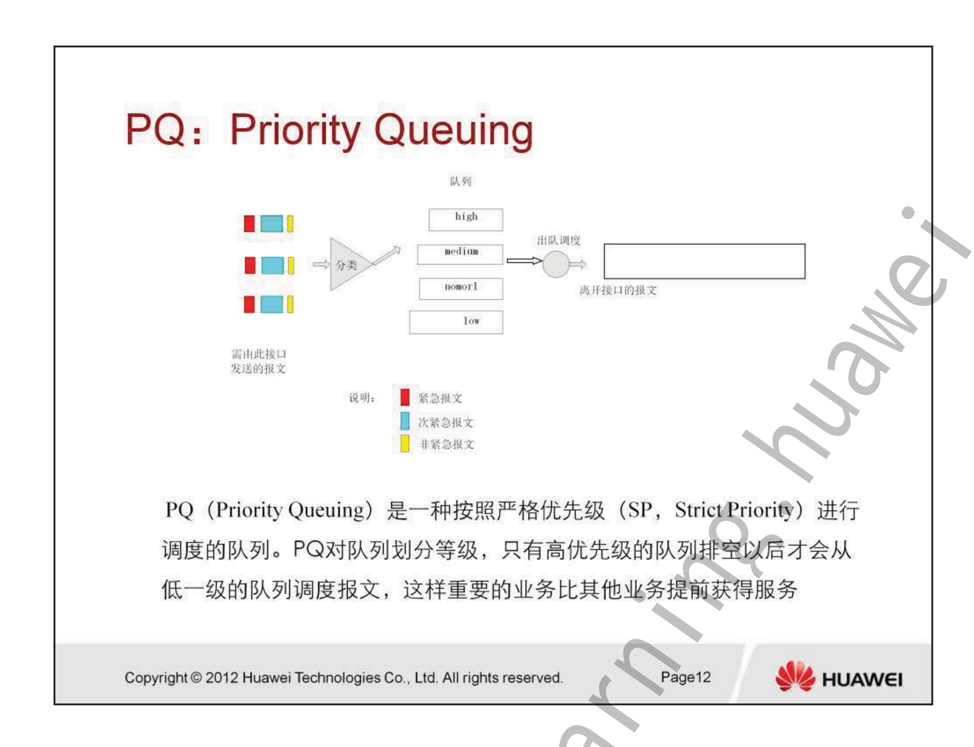
2、实现简单、复杂度低。

3、适合diffserv聚合后的端口。

缺点：

1、与RR调度算法一致，在报文长度不一致的时候，调度不准确。

2、在调度速率低的时候报文的时延控制的不好，时延抖动无法预期。



常规情况下PQ有四个级别的队列，分别为Top、Middle、Normal、Bottom，不过目前的设备大部分都实现了8个优先级队列，只要相对高的优先级队列有报文，就一直从高优先级队列取报文，所以PQ的优缺点是很明显的。

PQ队列的优点是可以保证高优先级队列的报文可以得到较大带宽、较低的时延、较小的抖动；缺点是低优先级队列的报文不能得到及时的调度，甚至得不到调度，即会出现“饿死”现象。

PQ具有如下特征：

1、可以使用ACL对报文进行分类，根据需要将报文入队列；

2、报文丢弃策略采用Tail Drop机制，且只有这一种机制；

3、队列长度可以设置为0，表示该队列无穷大，即进入该队列的报文不会被Tail Drop机制丢弃，除非内存耗尽了；

4、队列内部使用FIFO逻辑；

5、当从队列调度报文时，先从高优先级的队列调度报文。

从PQ特点可以看出，PQ保证某类流量尽可能得到最好的服务，而不管其它流量的“死活”。

优点:

1、高优先级队列的时延控制非常好。

2、实现简单，能够区分多种业务。

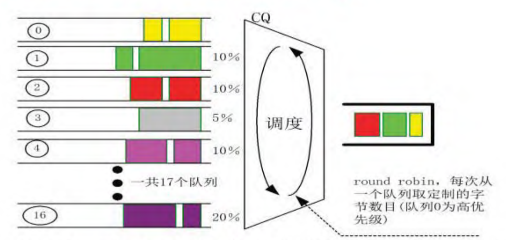
缺点:

1、无法做到带宽的合理分配，高优先级的流量比较大的时候，导致低优先级的流量“饿死（starvation）”。

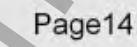
2、高优先级的时延得到保证的代价是牺牲低优先级的时延。

3、如果高优先级传送TCP流量，低优先级传送UDP流量，则TCP增加传送速率，导致UDP流量无法得到足够的带宽。

CQ：Custom Queuing



CQ（custom Queuing）可以支持17个队列，队列0用于系统队列，队列0和其他队列之间是SP的关系，只有队列0排空以后才能为其他队列提供服务，队列0一般用于协议报文。队列1至16没有优先级关系，采用轮询的方式，每次调度的时候从队列中调度固定字节数（预先配置），在轮询下一个队列之前，将数据包发出去。当某个队列已经调度了规定的字节数，或者该队列已经空，则轮询下一个队列。





这种队列需要在配置的时候为每一个队列指定每次调度的字节数量，每次队列调度的时候，如果报文长度超过配置的字节长度，则报文被调度，这种办法可以防止字节数配置的太小，导致某个队列阻塞。但是这种做法当配置字节数量比较小的时候，会导致带宽分配不准确。比如某个队列配置的调度字节数为500字节，而这个队列中的报文一般都是1000字节以上，则这个队列实际分配的带宽要比预期的大。如果将调度的字节数配置的比较大，则时延表现不好。CQ一次可以调度多个报文，数量为每次调度的字节数目能够容纳的包的个数。

优点：

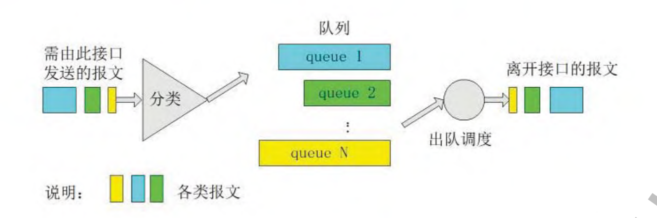
1、按照比例来分配带宽，当某个队列的流量小的时候，其他队列能等比的占用带宽。

2、实现简单。

缺点：

1、当配置字节数小的时候，带宽分配不准确，当配置字节数大的时候，时延抖动比较大。

WFQ：Weighted Fair Queuing



报文到达接口后，首先对报文进行分类，不同的流分入不同的队列。在出队的时候，WFQ按流的权重分配每个流应占的带宽。权重数值越小，所得的带宽越少。权重数值越大，所得的带宽越多。这样就保证了相同优先级业务之间的公平，体现了不同优先级业务之间的权值。

 Page15 HUAWEI

WFQ：加权公平队列（以后简称WFQ）对报文按流进行分类，对于IP网络，可以将相同源IP地址，目的IP地址，源端口号，目的端口号，协议号，IP优先级的报文属于同一个流，（有点象是同一个用户）而对于MPLS网络，具有相同的标签和EXP域值的报文属于同一个流。每一个流被分配到一个队列，尽量将不同的流分入不同的队列。在出队的时候，WFQ按权重来分配每个流应占有出口的带宽。权重的数值越小，所得的带宽越少。权重的数值越大，所得的带宽越多。这样就保证了相同优先级业务之间的公平，体现了不同优先级业务之间的权值。

例如：接口中当前有8个流，它们的权重值分别为1、2、3、4、5、6、7、8。则带宽的总配额将是所有权重值之和，即：1 + 2 + 3 + 4 + 5 + 6 + 7 + 8 = 36。

每个流所占带宽比例为：各自的权重/带宽的总配额。即，每个流可得的带宽比例分别为：1/36、2/36、3/36、4/36、5/36、6/36、7 /36、8/36。

由此可见，WFQ在保证公平的基础上对不同优先级的业务体现权值，而权值依赖于IP报文头中所携带的IP优先级。

优点:

1、按照字节粒度进行调度，调度公平。

2、能区分业务，分配权重。

3、时延控制的好，抖动小。

缺点:

1、实现复杂。

各种队列调度技术对比

队列技术	调度的时延/抖动（**在速率低的时候明显，速度绝对高的时候可忽略**）	公平性
FIFO	差	无
RR	差	依赖包长
WRR	差	依赖包长
PQ	高优先级队列的时延控制非常好	无
CQ	配置字节数小的时候，带宽分配不准确，当 配置字节数大的时候，时延抖动比较大差	一般
WFQ	时延控制较好，抖动小	好

 Page17 

各种队列调度技术的优缺点：

（一）FIFO：

优点：

1、简单。

缺点：

1、没有公平性，不同的流之间不互相隔离，当某一个流的带宽太大的时候会占用其他流的带宽，并且造成其他流的时延增加。

2、当拥塞发生的时候，FIFO对一部分报文进行丢弃。TCP的连接发现有丢包后，会降低传输的速度，来主动的避免拥塞，但是UDP是非连接的，不降低发送速率。导致FIFO中TCP和UDP的报文不平衡，TCP的流量太低。

3、一条流的突发流量可能占用全部buffer，将其他的流量全都阻断。

（二）RR：

优点：

1、隔离了不同的流，实现了队列之间对带宽的平等利用。

2、剩余带宽能够被其他队列平均分配。

缺点：

1、无法设置队列占用带宽的权重；

2、当不同队列中的报文长度不一的时候，调度不准确；

3、当调度速率低的时候，时延和抖动的问题比较突出，比如一个包到达一个空队列，而这个队列刚刚被调度完毕，则这个包要等到其他全部的队列调度完才能取得出接口的机会，这样会导致抖动比较大，但是如果调度速度非常高，则这种时延可以忽略，RR在高速路由器内部有很多应用。

（三）WRR：

优点：

1、能够按照权重来分配带宽，某个队列的剩余带宽能够为其他队列公平占用，低优先级的队列同样能够得到调度，不存在饥饿的问题。

2、实现简单、复杂度低。

3、适合diffserv聚合后的端口。

缺点：

1、与RR调度算法一致，在报文长度不一致的时候，调度不准确。

2、在调度速率低的时候报文的时延控制的不好，时延抖动无法预期。

（四）PQ

优点：

1、高优先级队列的时延控制非常好。

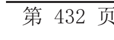
2、实现简单，能够区分多种业务。

缺点：

1、无法做到带宽的合理分配，高优先级的流量比较大的时候，导致低优先级的“饿死（starvation）”。

2、高优先级的时延得到保证的代价是牺牲低优先级的时延。

3、如果高优先级传送TCP流量，低优先级传送UDP流量，则TCP增加传送速率，导致UDP流量无法得到足够的带宽。



（五）CQ

优点：

1、按照比例来分配带宽，当某个队列的流量小的时候，其他队列能等比的占用带宽。

2、实现简单。

缺点：

1、当配置字节数小的时候，带宽分配不准确，当配置字节数大的时候，时延抖动比较大。

（六）WFQ

优点：

1、按照字节粒度进行调度，调度公平。

2、能区分业务，分配权重。

3、时延控制的好，抖动小。

缺点：

1、实现复杂。

队列技术在产品中的实现

目前产品中，主要使用FIFO、WFQ、PQ三种队列技术来实现拥塞管理。

对于队列配置，用户无须关心采用什么抽象的调度算法，只需关心队列所承载业务的外在流量参数特征，比如保证多少兆的带宽、峰值最多多少兆的带宽、要占剩余带宽的比例权重等。根据配置的流量参数选用不同的调度算法来严格保证用户的配置。

端口队列调度采用PQ+WFQ调度算法，采用这种调度优势在于，既能使时延敏感的实时业务得到保证，也使优先业务的报文的带宽占用可以绝对优先，又可以为不同优先级的流根据配置的权重分配不同的带宽。

　Page20



对于DiffServ模型，系统为每个端口预留8个业务队列，分别对应BE，AF1至AF4，EF，CS6，CS7等业务类别，对AF1~AF4以及BE队列默认配置成WFQ调度，根据配置的权重参数按比例分配带宽。EF，CS6，CS7队列默认配置PQ调度，这种按照绝对优先级调度，一般是时延敏感的业务采用PQ调度。


目 录
拥塞管理
拥塞避免
Copyright © 2012 Huawei Technologies Co., Ltd. All rights reserved.
Page21
HUAWEI

拥塞避免机制

拥塞避免是一种流控机制，它可以通过监视网络资源（如队列或内存缓冲区）的使用情况，在拥塞有加剧的趋势时，主动丢弃报文，通过调整网络的流量来解除网络过载。

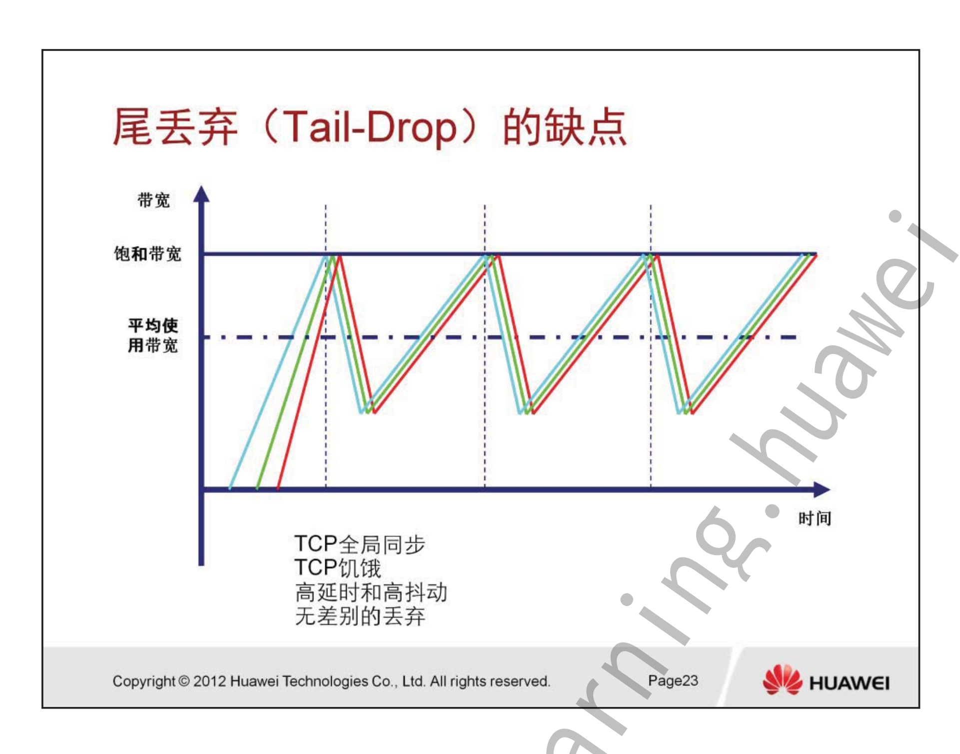
传统的丢包策略采用尾部丢弃（Tail-Drop）的方法。当队列的长度达到某一最大值后，所有新到来的报文都将被丢弃。这种丢弃策略会引发TCP全局同步现象。

为避免TCP全局同步现象，可使用RED或WRED。

- RED：Random Early Detection，随机早期检测
- WRED：Weighted Random Early Detection，加权随机早期检测

Page22





1、TCP全局同步

拥塞避免机制传统的处理方法是尾丢弃，当队列的长度达到规定的最大长度时，所有到来的报文都被丢弃。对于TCP报文，由于大量的报文被丢弃，将造成TCP超时，从而引发TCP的慢启动和拥塞避免机制，使TCP减少报文的发送。当队列同时丢弃多个TCP连接的报文时，将造成多个TCP连接同时进入慢启动和拥塞避免，称之为TCP全局同步。这样多个TCP连接发向队列的报文将同时减少，使得发向队列的报文的量不及接口发送的速度，减少了链路带宽的利用。并且，发向队列的报文的流量总是忽大忽小。使线路的上的流量总在极少和饱满之间波动。另外也会影响特定流量的延时和抖动。

2、 TCP饥饿

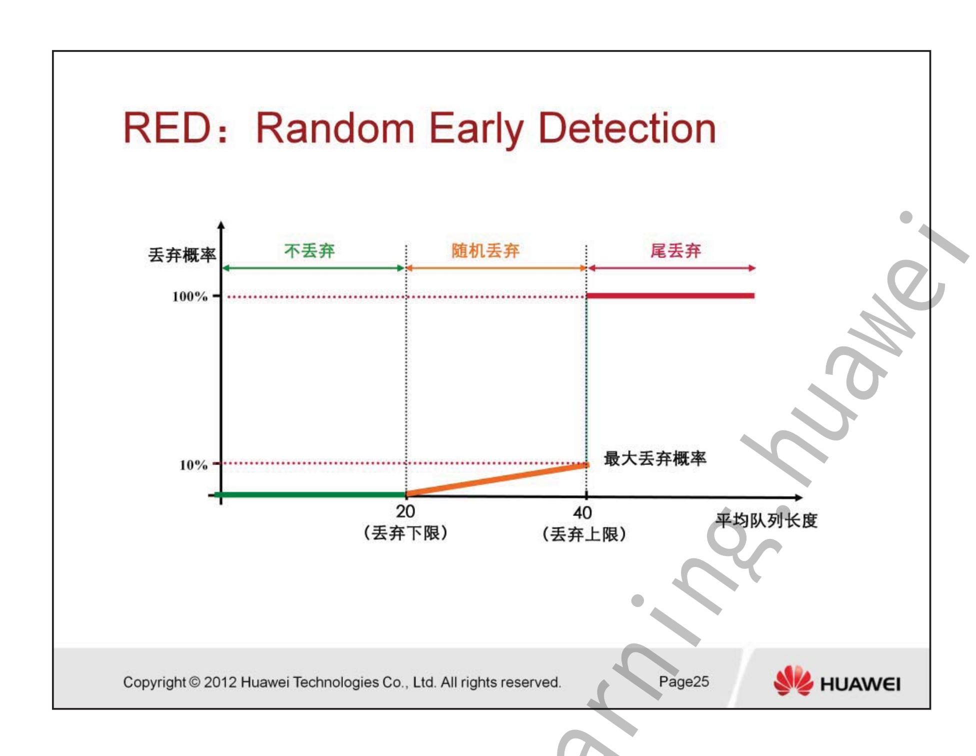
尾丢弃会造成TCP流量之间分配带宽不均衡，一些“贪婪”的流量会占用大部分的带宽，而普通的TCP流量分配不了带宽而“饿死”。特别是网络中既有TCP又有UDP流量的时候，TCP流量因为窗口机制（尾丢弃造成滑动窗口减小）而释放带宽，UDP流量没有窗口机制，于是UDP流量会迅速占用TCP释放的带宽，最终

造成UDP流量占用了所有带宽而TCP流量因没有带宽分配而“饿死”。

3、高延时和高抖动

拥塞导致延时和抖动增加。

4、无差别的丢弃，没有区分各种不同优先级的报文。



为了避免尾丢弃的这些问题，必须在端口将要拥塞之前进行丢弃。RED就是一种在队列拥塞之前进行报文丢弃的一种拥塞避免机制。RED会主动丢弃可能造成拥塞的报文。他能够使TCP会话所占用的输出带宽缓慢的降低，不会引起大量的TCP全局同步以及TCP饥饿，还能够降低平均队列长度。

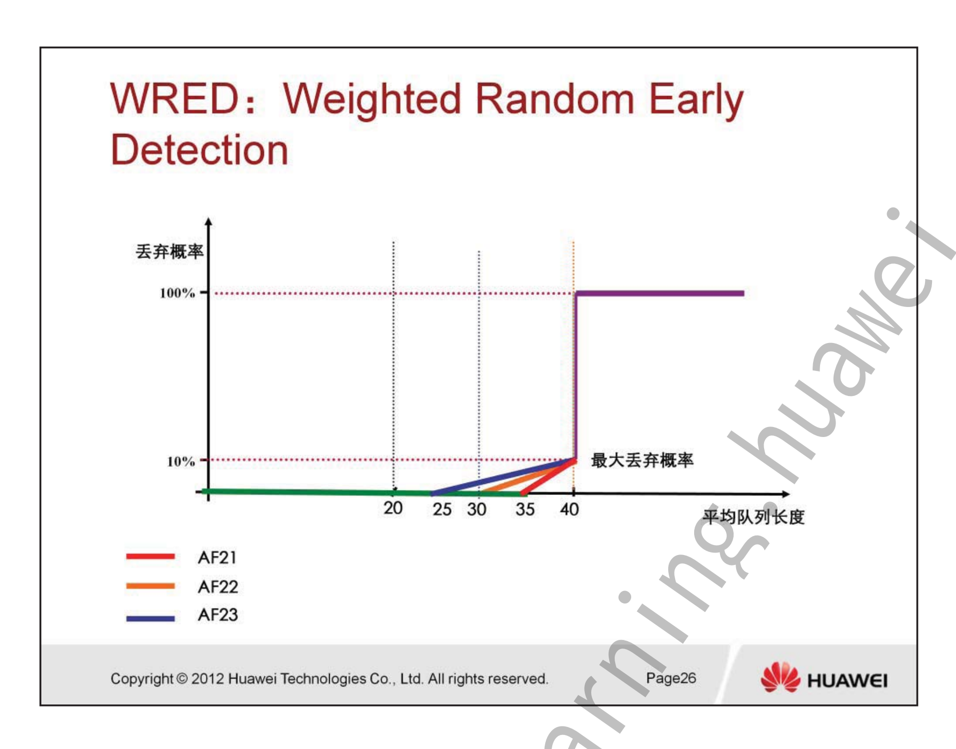
RED共有三种丢弃模式：绿色报文不丢弃、黄色报文概率丢弃、红色报文全丢弃。

这三种模式是以队列丢弃的上下两个阀值（low-limit和high-limit）所决定的。

1、绿色报文——当平均队列长度小于low-limit时被标记为绿色，不进行丢弃。

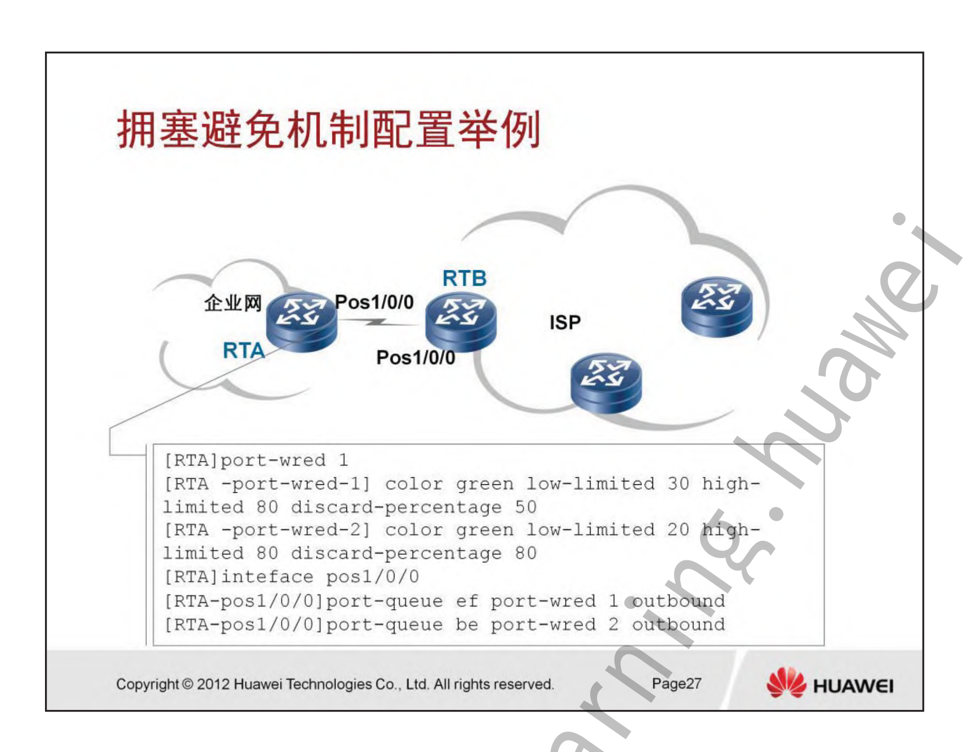
2、黄色报文——当平均队列长度介于low-limit和high-limit之间时，报文被标记为黄色，进行概率丢弃。并且，队列的长度越长，丢弃的概率越高。

3、红色报文——当平均队列的长度大于high-limit时，报文被标记为红色。并且进行全部丢弃。（此时进行的是尾丢弃！）



WRED（Weighted Random Early Detection）与RED的区别是引入了优先权，不同的优先权可以有不同的丢弃策略。每一个丢弃策略都包含有RED的三个参数：下限阀值、上限阀值以及最大丢弃概率。目前WRED优先权可以根据DSCP和IP优先级进行划分，对低优先级报文的丢弃概率大于高优先级的报文。

DSCP AF PHB表示为aaadd0，其中‘aaa’表示流量类别，‘dd’表示丢弃优先级。比如AF21(010010)、AF22(010100)和AF23(010110)属于同一个流量类，并且在拥塞发生的时候，丢弃可能性AF21<AF22<AF23，所以在配置WRED参数时，可以按图中所示进行配置。对标记为AF21的流量下限设为35，上限设为40；标记为AF22的流量下限设为30，上限设为40；标记为AF23的流量下限设为25，上限设为40。另外，在达到上限时的丢弃概率是10%。因此，在可能发生拥塞之前，AF23的数据包最有可能最先被丢弃。



在流量整形部分对RTA的POS1/0/0端口配置了两个服务类EF和BE的调度算法分别为PQ和WFQ，这里针对EF服务类配置了WRED模板1，BE服务类配置了模板2。具体的配置步骤如下：

1、对优先级较高的ef流将低门限和高门限值配置较高，丢弃概率设置较小，缓存较多的报文。

模板1 color green low-limited 30 high-limited 80 discard-percent 50

2、 在接口下指定EF服务类指定WRED模板：

[RTA-pos1/0/0]port-queue ef port-wred 1 outbound

3、对优先级较低的be流将低门限和高门限值配置较低，丢弃概率可以设置比较大，缓存较少的报文。

模板2 color green low-limited 20 high-limited 80 discard-percent 80

4、在接口下指定BE服务类指定WRED模板：

[RTA-pos1/0/0]port-queue be port-wred 2 outbound

命令解释如下：

port-wred port-wred-name命令用来创建类队列WRED对象。

color { green | yellow | red } low-limit low-limit-percentage high-limit high-limit-percentage discard-percentage discard-percentage

命令用来配置WRED模板的拥塞避免机制，每个模板最多支持3种颜色报文的处理，包括green、yellow和red。每个WRED模板最多支持红、黄、绿3种颜色报文的处理。一般绿色报文设置的丢弃概率比较小，高、低门限值比较大；黄色报文次之；红色报文设置的丢弃概率最大，高、低门限值最小。

通过配置WRED模板，用户可以为队列设定阈值(包括高阈值、低阈值)和丢弃概率。

当队列的长度小于低阈值时，不丢弃报文；当队列的长度在低阈值和高阈值之间时，WRED开始随机丢弃报文(队列的长度越长，丢弃的概率越高)；当队列的长度大于高阈值时，丢弃所有的报文。每种颜色报文的门限值和丢弃概率都是可配置的，实现灵活。

针对带宽比较大的队列,一般说明这种流的优先级比较高，可以配置WRED的低门限和高门限值都比较高，丢弃概率可以设置比较小，缓存较多的报文，带宽小的队列设置反之。

注意:

1、WRED模板只是配置了某种颜色对应的丢弃概率，并未指定对应何种服务类。在实际使用时需要为不同服务级别的队列在出方向指定WRED模板。

2、目前在接口配置WRED时，对于上行的WRED没有作用。只是对下行的队列有作用。命令各参数解释如下:

green，yellow，red：根据DSCP值中的后面两位确定的颜色值。

low-limit-percentage：WRED丢弃的低门限百分比，表示WRED丢弃的低门限值占类队列长度的百分比。整数形式，取值范围是0～100，单位是percentage，缺省值为100。

high-limit-percentage：WRED丢弃的高门限百分比，表示WRED丢弃的高门限值占类队列长度的百分比。整数形式，取值范围是low-limit-percentage～100，单位是percentage，缺省值为100。

discard-percentage-value：WRED的丢弃概率百分比。整数形式，取值范围是1～100，缺省值为100。



问 题

什么是拥塞管理?

常用的队列调度技术有哪些?

什么是拥塞避免?

常用的拥塞避免有哪些?

 Page29



答案:

- 拥塞管理是指网络在发生拥塞时，如何进行管理和控制。
- 常见的队列调度技术有以下几种：FIFO、RR、WRR、PQ、CQ、WFQ。
- 拥塞避免是一种流控机制，它可以通过监视网络资源（如队列或内存缓冲区）的使用情况，在拥塞有加剧的趋势时，主动丢弃报文，通过调整网络的流量来解除网络过载。
- 常用的拥塞避免有尾丢弃、RED、WRED。



谢谢

www.huawei.com


链路效率机制
www.huawei.com
Copyright © 2012 Huawei Technologies Co., Ltd. All rights reserved.
HUAWEI



培训目标

学完本课程后，您应该能：

- 理解链路效率机制的原理。
- 掌握链路效率机制的配置。

 Page1 HUAWEI

链路效率机制

IPHC：IP Header Compression，IP报文头压缩

- RTP报文头压缩
- TCP报文头压缩

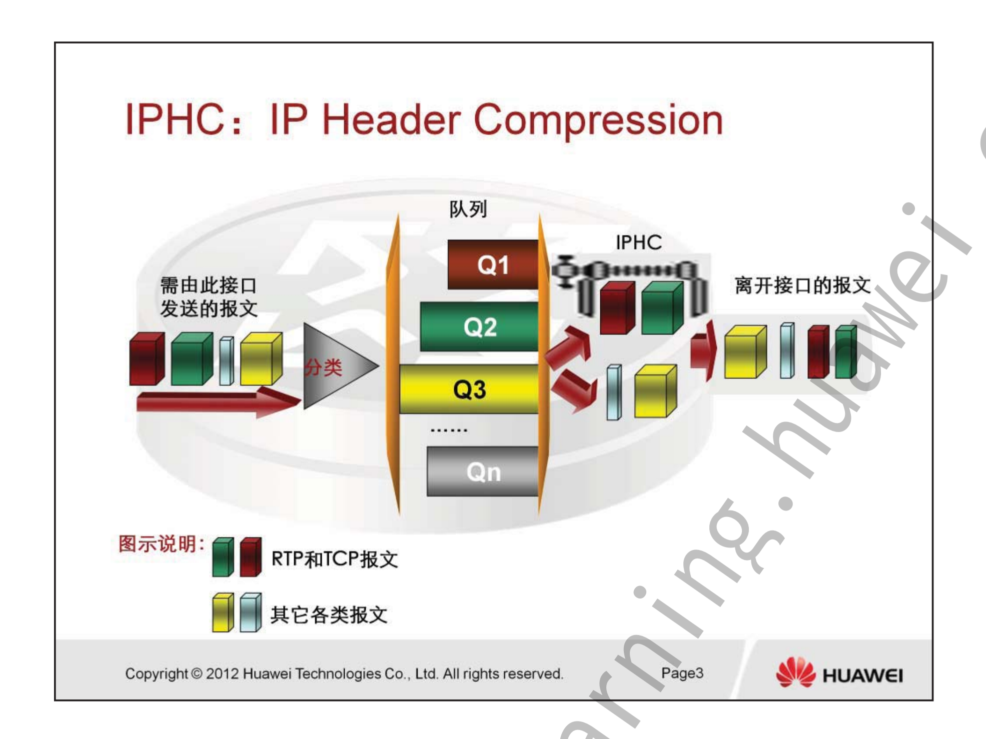
LFI：Link Fragmentation and Interleaving，链路分片与交叉

 Page2 

VRP提供了两种链路效率机制：IP报文头压缩协议（IP Header Compression，IPHC）和链路分片与交叉（Link Fragmentation and Interleaving，LFI）。其中IP报文头压缩协议可以对RTP和TCP报文头进行压缩。

对于同一个流的数据部，IP头部的大部分字段是相同的，因此可以对这些字段进行压缩，提高链路传输的效率。

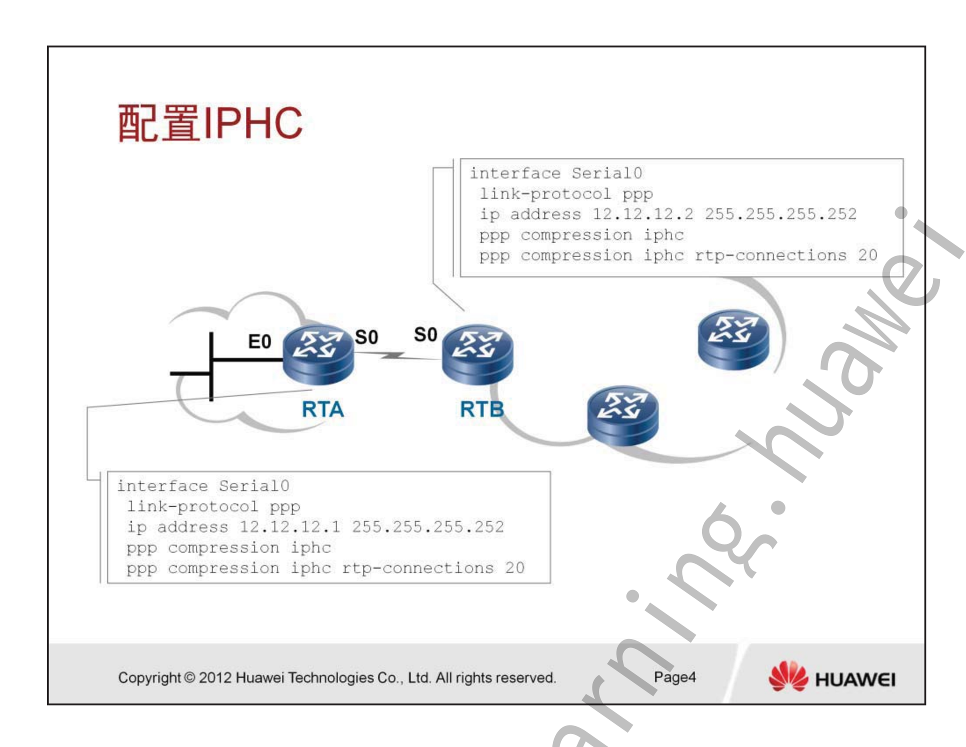
LFI技术主要在低速链路上使用，目的是减小实时数据报文的延时和抖动。



IP报文头压缩协议（IP Header Compression，IPHC）是一个主机-主机协议，用于在IP网络上承载语音、视频等实时多媒体业务，是在PPP链路和FR链路应用的低速链路技术。IPHC支持对RTP和TCP报文头的压缩。

RTP包括数据部分和头部分，RTP的数据部分相对较小，而RTP的头部分较大。12字节的RTP头，加上20字节的IP头和8字节的UDP头，就是40字节的IP/UDP/RTP头。而RTP典型的负载是20字节到160字节。为了避免不必要的带宽消耗，可以使用IPHC特性对报文头进行压缩。IPHC将IP/UDP/RTP头从40字节压缩到2～4字节，对于40字节的负载，头压缩到4字节，压缩比为（40+40）/（40+4），约为1.82，可见效果是相当可观的。

对于TCP数据包，IP头加上TCP头一共40字节，使用TCP压缩，可以压缩到3～5字节。



示例中在串行链路的两端执行IP头压缩，其中最大RTP和TCP头压缩的连接数为20。

具体配置命令解释如下。

启动IP头压缩

1、进入系统视图：system-view

2、进入接口视图：interface interface-type interface-number

3、启动IP头压缩：ppp compression iphc [nonstandard]

ppp compression iphc命令用来启动某接口上的IP头压缩，undo ppp compression iphc命令用来关闭IP头压缩功能。

缺省情况下，不启动接口上的IP头压缩功能。

IP报文头压缩命令将启动IP/UDP/RTP头压缩，以及建立RTP会话的IP/TCP头压缩。

用户必须在链路的两端同时配置IP头压缩命令。

在配置完成后，只有对接口进行shutdown与undo shutdown操作后，配置才能生效。如果是应用在MP上，要对所有MP捆绑的接口执行shutdown与undoshutdown操作。

配置RTP头压缩的最大连接数。

1、进入系统视图：system-view

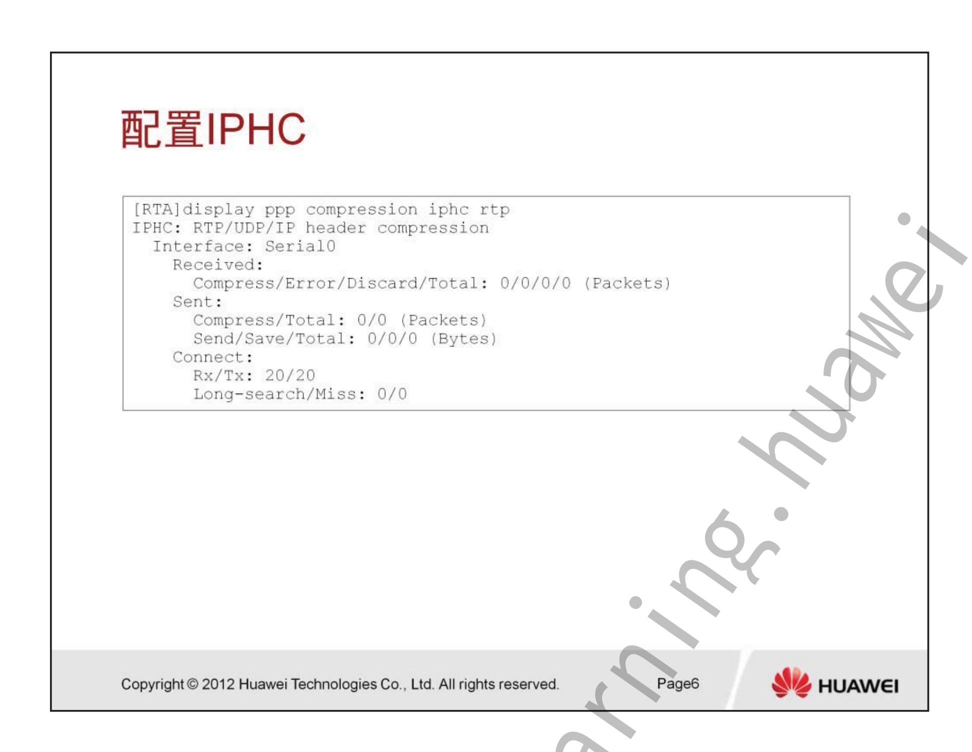
2、进入接口视图：interface interface-type interface-number

3、配置RTP头压缩的最大连接数：ppp compression iphc rtp-connections number

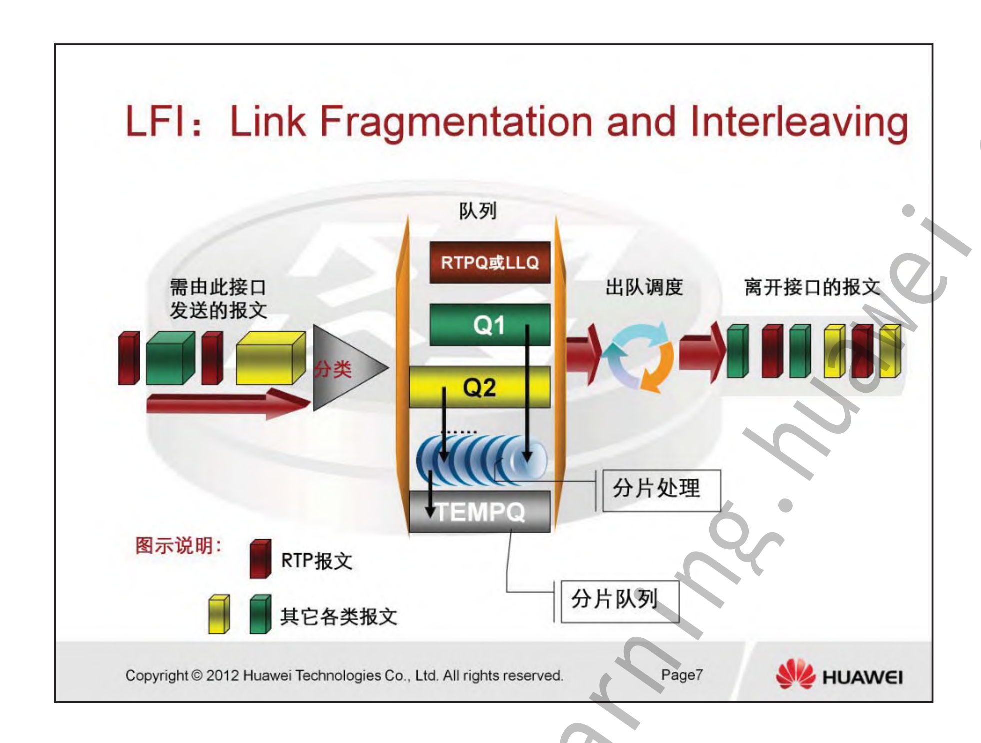
number：指该接口上IPHC功能的RTP头压缩的最大连接数，取值范围3～1000，缺省值为16。

ppp compression iphc rtp-connections命令用来指定某一个接口上允许存在RTP头压缩连接的总数，undo ppp compression iphc rtp-connections命令可以取消配置，恢复缺省值。

配置将在对接口进行shutdown与undo shutdown操作后生效。如果是对MP进行配置，则shutdown与undo shutdown操作必须在所有MP上实施。



配置完毕可以使用 “display ppp compression iphc rtp”命令观察IPHC的效果。



链路分片与交叉是在PPP链路和FR链路应用的低速链路技术。

在低速串行链路上，实时交互式通信，如Telnet和VoIP，往往会由于大型分组的发送而导致阻塞延迟，例如，正好在大报文被调度而等待发送时，语音报文到达，它需要等该大报文被传输完毕后才能被调度。对于诸如交互式语音等实时应用而言，大报文导致的这种阻塞延迟太长了，对端将听到断断续续的话音。交互式语音要求端到端的延迟不大于100～150ms。

一个1500bytes（即通常MTU的大小）的报文需要花费215ms穿过56Kbps的链路，这超过了人所能忍受的延迟限制。为了在相对低速的链路上限制实时报文的延迟时间，例如56Kbps Frame Relay或64Kbps ISDN B通道，需要一种方法将大报文进行分片，将小报文和大报文的分片一起加入到队列。

LFI将大型数据帧分割成小型帧，与其他小片的报文一起发送，从而减少在速度较慢的链路上的延迟和抖动。

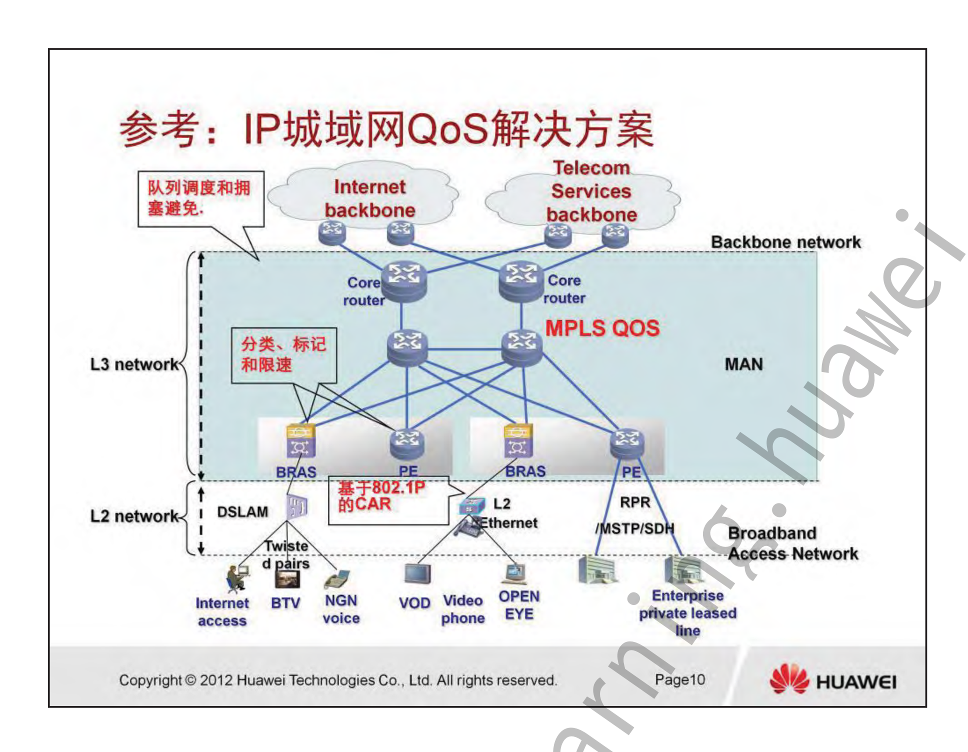
上图描述了链路分片与交叉的处理过程。大报文和实时报文一起到达某个接口时，除了RTP实时队列和LLQ队列中的报文外，其他队列中的大报文将被分成若干小包放入分片队列进行发送；但

如果此时RTP实时队列和LLQ中有缓存的报文，则优先调度RTP实时队列和LLQ队列，否则继续调度分片队列，这样就避免了在低速链路上传送大包对实时报文造成的时延与抖动。



Q：常用链路效率机制有哪些

A：IPHC、LFI。



随着网络业务的不断升级，IP网所承载的业务已经由单纯的互联网业务发展到了数据、语音和视频等综合业务，流媒体、IPTV等新业务对运营商本地IP网QoS的要求从简单的网络可靠性发展到了时延、时延抖动、丢包率及网络的可靠性等全面的QoS要求。

QoS技术的主要目的是在链路拥塞的情况下，仍能够保证重要业务不受影响。其中主要的参数为网络时延及丢包率。目前应用较多的IP QoS技术是队列技术及WRED技术。QoS并没有创造带宽，只是根据应用程序的需求以及网络状况来管理带宽。

通常运营商网络由宽带接入网、城域网和骨干网三部分组成。宽带接入网一般是由DSLAM或二层交换机组成的二层网络，在QoS方面可以部署802.1P区分不同用户的优先级，并配置基于802.1P的CAR来限制入网流量。

城域网BRAS或PE设备上可以通过进一步流分类区分出同一用户的不同业务流，为不同的业务流选择合适的队列调度算法和拥塞避免方法，从而为不同的业务提供不同的服务保证。

QoS的应用随着网络需求的变化而不断变化， QoS是一项复杂的工程，其中不仅仅包括这里所提到队列，拥塞避免技术。在某些

更为复杂的网络，还可以利用灵活QinQ技术实现基于用户基于业务的Qos技术以及分层级的HQoS技术。对这部分内容感兴趣的读者可以参考一些其他技术类的文档。

谢谢

www.huawei.com

在线学习资料支持

您可以在华为企业业务网站获得E-Learning课程、培训教材、产品资料、软件工具、技术案例等:

1、 E-Learning课程：登录*华为在线学习网站*，进入 “*华为培训/在线学习*” 栏目

免费E-Learning课：对网站所有用户免费开放

职业认证E-Learning课：通过任何一项职业认证即可学习所有职业认证培训E-Learning课程

渠道赋能E-Learning课：对华为企业业务合作伙伴免费开放

2、培训教材：登录*华为在线学习网站*，进入 “*华为培训/面授培训*” ，在具体课程页面即可下载教材。

华为职业认证培训教材、华为产品技术培训教材。无需注册即可下载

3、华为在线公开课(LVC)：http://support.huawei.com/ecommunity/bbs/10154479.html

企业网络、UC&C、安全、存储等诸多领域的职业认证课程，华为讲师公开授课

4、产品资料下载：http://support.huawei.com/enterprise/#tabname=productsupport

5、软件工具下载：http://support.huawei.com/enterprise/#tabname=softwaredownload

更多内容请访问：

- http://learning.huawei.com/cn
- http://support.huawei.com/enterprise/
- http://support.huawei.com/ecommunity/